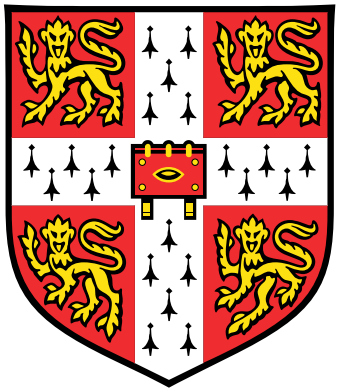
The genomic basis of species barriers in *Heliconius* butterflies



Ana Leonor Pessoa Pinharanda

Department of Zoology

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

St John's College December 2017

Para as três mulheres da minha vida

Para a minha família and friends

Pour toi

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit of 60,000 words for the Degree Committee in Biology.

Ana L P Pinharanda

September 2017

ACKNOWLEDGMENTS

This work would have not been possible without the supervision of Prof Chris D Jiggins. Chris was enthusiastic, present and supportive from the moment we met in 2012 at PopGroup all throughout my PhD. I was given the guidance a PhD student needs but also the freedom to develop my own project. It feels really special to have had a number of amazing opportunities as part of my PhD: from insightful discussions, to the fieldwork season in Panama or the bioinformatics course in Leipzig, thank you.

Dr John Davey is another person without whose support this work would have not been possible. John spent innumerable hours teaching me how to approach and solve a scientific problem. I have learnt from him how rewarding questioning the seemingly obvious can be. Being co-supervised by John has helped me to be become a better colleague and scientist and I am grateful for that.

I was also extremely lucky to be part of the Butterfly Genetics Group. During the years I have had amazing colleagues with whom I had the pleasure to both work and socialise with. I am especially grateful to Dr Simon Martin and Dr Camille Roux who discussed population genetic concepts with me and helped me tackle some practical bioinformatics problems. I am thankful to Dr Richard Merrill with whom I had lengthy and insightful scientific discussions and without whom my time in Panama would have not been the same – living with him and Dr Denise Dalbosco Dell'Aglio in Gamboa was really special. Thank you Gabriela Montejo-Kovacevich, Kathy Darragh, Sarah Barker, Dr Ian Warren, Dr Richard Wallbank, Dr Markus Möest, Dr Steven Van Belleghem, Laura Hebberecht López and Dr Pasi Rastas for making my time as a PhD student in the Butterfly Genetics group a truly fun learning experience. Joe Hanly and I started our PhDs in the same day and I cannot

imagine these years without him – he has been a great colleague and an incredible friend.

During my PhD I have spent seven months in Panama. The Butterfly Group at the Smithsonian Tropical Research Institute in Gamboa, has taught me how to work and care for *Heliconius*. I cannot thank them enough for cheering with me when a mating occurred, or comfort me when geckos ate my butterflies. I specially thank Adriana Tapia, Oscar Paneso, Moises Abanto, Tim Thurman, Liz Evans and Rachel Crisp for their every day positivity and dedication to insectary life. Finally, an immense thank you to Dr Owen McMillan. Owen always made sure we had all the resources we needed to do research, and was available for helpful scientific discussions.

I would also like to thank Dr Andrea Manica, Dr John Welsh and Dr Steven Montgomery for helpful advice throughout my PhD. The 4-week course in Leipzig organized my Dr Katja Nowick and Dr Rui Faria introduced me to the bioinformatic tools I would use throughout my project and it was vital for my progress. Finally, I owe a special thanks to Jenny Barna and Stuart Rankin from the Darwin High Performance Computing Cluster who have helped me numerous time with software version and installation issues.

I gratefully acknowledge the funding: National Environment Research Council, St John's College 10th Term Funding, a grant from the Cambridge Philosophical Society and a grant for financial hardship from the Elliot and Leathersellers' Company Funds.

* * *

I would also like to thank all my friends and family who helped me believe I had what it took to do a PhD. There are not enough words to thank my Mother, my Grandmothers, my Aunt, my Dad, my two sisters Beatriz and Francisca, and the rest of my family – I am deeply grateful for their unconditional support and encouragement. Since the days my Mum allowed

me to rear *Bombyx* in a large cardboard box and my Grandmother picked mulberry tree leaves from the side of a dual carriageway, I knew loved Lepidoptera.

Thank you to all my friends from undergraduate at The University Manchester for throwing the best parties in London that kept me going during the process. Thank you to all my Cambridge friends for the live music and dancing that made me smile. Especially thanks to the Madingley girls for making my time in Cambridge feel like home. To Jérémie Le Pen, thank you for being there everyday and every step of the way.

SUMMARY

Understanding the genetics underlying the speciation process has been a long-standing goal of evolutionary biology. Studying inter-population crosses can elucidate the genetic architecture of reproductive isolation and, ultimately, the process of speciation. Hybridization between two species is often maladaptive and results in offspring with decreased fitness compared to the parental forms. Recently, with the development of molecular and genomic tools, it has become possible to understand how and when reproductive isolation arises and what are the underlying mechanisms in the evolution of genetic incompatibilities.

Heliconius is a genus of neotropical butterfly best know for their Müllerian mimicry. Here I focus on *Heliconius cydno* and *Heliconius melpomene*, two hybridising sympatric species with low levels of inter-specific hybridisation that nonetheless results in genome-wide signatures of admixture. I show that hybrids develop ovarian tissue and, occasionally, oocytes; and use genomic approaches to examine several potential mechanisms underlying post-zygotic isolation between *H. cydno* and *H. melpomene*.

Firstly, I investigate evolution by gene duplication and identify loci putatively under divergent selection that may play a role in species divergence and speciation. Secondly, I quantify sexually dimorphic expression in *H. melpomene*, and calculate rates of molecular evolution between *H. melpomene and H. erato*. Thirdly, I identify differentially expressed genes in the *H. cydno* x *H. melpomene* F1 hybrids that may be involved in the species barrier. Finally, investigate whether epigenetic silencing mechanisms could underlie post-zygotic isolation between *H. cydno* and *H. melpomene* by quantifying transposable element expression and small RNAs.

Overall, I identify loci that merit further investigation for their potential in

maintaining reproductive barriers between these two species. I show that different regions of the genome evolve at different molecular rates but there is no faster-Z effect, and consider how might this affect evolution of reproductive isolation. Finally, I show that aberrant epigenetic silencing, a mechanism behind hybrid sterility that is common in other species, is not correlated with post-zygotic isolation between *H. cydno* and *H. melpomene*.

322 words

COLLABORATIONS AND PUBLICATIONS

Chapter 1, The comparative landscape of duplication in *Heliconius melpomene* and *Heliconius cydno,* has been published as (Appendix A):

> **Pinharanda A**, Martin SH, Barker SL, Davey JW, Jiggins CD (2017). The comparative landscape of duplications in Heliconius melpomene and Heliconius cydno. *Heredity* 118: 78–87.

I am thankful to Dr Simon Martin for the high-throughput *Heliconius melpomene rosina* DNA sequencing data, read mapping and SNP calling used in Chapter 2, Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*. I am also grateful to Marjolaine Rousselle and Dr Nicolas Galtier from Université de Montpellier for running scripts for molecular evolution rates using the *Modelling Approach* and for general discussions and advice. Finally, I am thankful to Dr Sujai Kumar from LepBase who ran a pipeline to improve the completeness of the *H. melpomene* annotation.

In Chapter 3, Sterility in *Heliconius cydno* x *Heliconius melpomene* F1 female hybrids: a phenotypic and gene expression study of hybrid incompatibles, I am thankful to Dr John Davey who made the *H. cydno* reference genome transfer from *H. melpomene*.

In Chapter 4, piRNA mediated epigenetic silencing does not underlie post-zygotic isolation between *Heliconius cydno* and *Heliconius melpomene*, I am thankful for Dr Jérémie Le Pen from the University of Cambridge for help with the sRNA extractions and for general discussions and advice. I am also grateful for Dr Sam Lewis and Dr Frank Jiggins from the University of Cambridge for running the repeat element annotation pipeline, sharing sRNA analysis scripts and providing guidance on how to best develop the project.

OTHER PROJECTS

Some aspects of my work are not included in this thesis but have been published as part of larger collaborations.

1) Contributed to the collective writing of sections on the genomic landscape of speciation

Merrill RM, Dasmahapatra KK, Davey JW, Dell'Aglio DD, Hanly JJ, Huber B, Jiggins CD, Joron M, Kozak KM, Llaurens V, Martin SH, Montgomery SH, Morris J, Nadeau NJ, **Pinharanda A** *et al.* (2015). The diversification of Heliconius butterflies: what have we learned in 150 years? *Journal of Evolutionary Biology* **28**: 1417–1438.

2) Performed the lab work that validated the putative *de novo* mutations

Keightley PD, **Pinharanda A**, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, *et al.* (2015). Estimation of the spontaneous mutation rate in Heliconius melpomene. *Molecular Biology and Evolution* **32**: 239–243.

3) Performed the *H. cydno* crosses and DNA extractions

Davey JW, Barker SL, Rastas PM, **Pinharanda A**, Martin SH, Durbin R, *et al.* (2017). No evidence for maintenance of a sympatric Heliconius species barrier by chromosomal inversions. *Evolution Letters*: n/a–n/a.

ABBREVIATIONS

BDM	->	Bateson-Dobzhansky-Muller
BDMI	->	Bateson-Dobzhansky-Muller incompatibilities
BWA	->	Burrows-Wheeler Aligner
CPCP	->	*H. cydno* Panama
CPMP	->	*H. cydno* x *H. melpomene* Panama
FDR	->	False discovery rate
HMEL	->	*H. melpomene* sample
HMM	->	Hidden Markov Model
LINE	->	Long interspersed nuclear elements
LTR	->	Long terminal repeats
MACSE	->	Multiple alignment of coding sequences
MPMP	->	*H. melpomene* Panama
NGS	->	Next generation sequencing
PCA	->	Principal component analysis
PCR	->	Polymerase chain reaction
PE	->	Pair end reads
QTL	->	Quantitative trait loci
SINE	->	Short interspersed nuclear elements
SNP	->	Single nucleotide polymorphism
TE	->	Transposable element
TF	->	Transcription factor
UTR	->	Untranslated region
VCF	->	Variant call format
WG	->	Whole-genome
WGS	->	Whole-genome sequence(ing)

TABLE OF CONTENTS

INTRODUCTION

The architecture of reproductive isolation: how studying inter-specific crosses can elucidate the process of speciation

CHAPTER 1

The comparative landscape of duplication in *Heliconius melpomene* and *Heliconius cydno*

CHAPTER 2

Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*

CHAPTER 3

Sterility in *Heliconius cydno* x *Heliconius melpomene* F1 female hybrids: a phenotypic and gene expression study of hybrid incompatibles

CHAPTER 4

piRNA mediated epigenetic silencing does not underlie post-zygotic isolation between *Heliconius cydno* and *Heliconius melpomene*

Appendix C

Appendix D

The genetic architecture of reproductive isolation: how studying inter-specific crosses can elucidate the process of speciation

Introduction

Evolution is the change in genotype frequencies through time and, while genotype frequencies are easily measured, their change is not. Four population genetic forces govern these changes: mutation, recombination, drift, and natural selection. Selection, first articulated by Darwin, plays a central role in the evolution of complex phenotypic traits (Lenski *et al.*, 2003). Mutation, recombination and drift are non-adaptive in the sense they are not a function of the fitness properties of individuals. Mutation, including insertions, deletions and duplications, is the ultimate source of genetic variation on which selection acts (Lynch, 2010). Recombination, including crossing-over and gene conversion, sorts genetic variation (Barton, 1995). Genetic drift is independent of the other three forces, and results in random changes of genotype frequencies due to offspring number and allele segregation patterns (Charlesworth, 2009). Finally, epistatic effects also determine evolvability, and incorporating these interactions onto the existing evolution population genetic framework is an active area of research (Carter *et al.*, 2005). The relative strength, direction and variation of these forces over time, determines how adaptive and non-adaptive processes tailor genomic architecture, and shapes the way evolution proceeds.

An understanding of molecular biology, combined with studies of molecular variation within and between species, can result in reliable estimates of the relative strength of the different evolutionary forces operating at the genomic level. Specifically, addressing the genomic basis of species differences advances our understanding of the genetics of speciation and may allow to determine the genetic basis of species origins. Hybridization may slow differentiation through gene flow and recombination; or accelerate speciation via adaptive introgression or allopolyploidization. Therefore, in cases for which inter-specific hybrids exist, studying them can reveal the genetic basis of barriers to inter-specific gene flow and help to determine the relative contribution of adaptive divergence through the speciation process (Coyne and Orr, 2004). It is important to note, however, that the genetics of speciation is not restricted to the identification of hybrid sterility or inviability genes. Any genomic loci that drives ecological, sexual, pre- or post-zygotic isolation is a barrier to inter-specific gene flow (Wolf *et al.*, 2010).

In the following chapter, I highlight how studying the different evolutionary forces is key to understand the underlying mechanisms in the evolution of genetic incompatibilities. I consider how genomic architecture might shape such incompatibilities, and introduce *Heliconius* as a system to study the genetic basis of barriers to inter-specific gene flow. I describe how DNA-level features can differ between autosomes and sex chromosomes and how this affects the evolution of reproductive isolation. I consider how mobile genetic elements (a major source of genomic mutations) and gene duplications (a class of structural variations) contribute of genome expansion, and often also to reproductive isolation. Finally, I discuss studies of gene expression between different populations or species and their hybrids to illustrate how variation at coding and non-coding regions of the genome can impact reproductive fitness.

Heliconius in evolutionary biology research

Heliconius are a group of Neotropical butterflies that have contributed to answering a broad range of evolutionary questions from taxonomy to behaviour. Historically, studies of the genetics of *Heliconius* butterflies have focused on loci controlling colour patterns, with many races diverging at these loci alone (Nadeau *et al.*, 2012; Martin *et al.*, 2013). Throughout the 19th century the first evolutionists were drawn to the taxa's wing-pattern mimicry after observing that divergent lineages have repeatedly converged on virtually identical wing warning patterns (Figure 1). Henry Walter Bates developed mimicry theory after observing *Heliconius* butterfly patterns (Bates, 1862). Bates interpreted the differences in colour pattern between separate geographic populations as support for Darwin's theory of species mutability (Darwin, 1859).

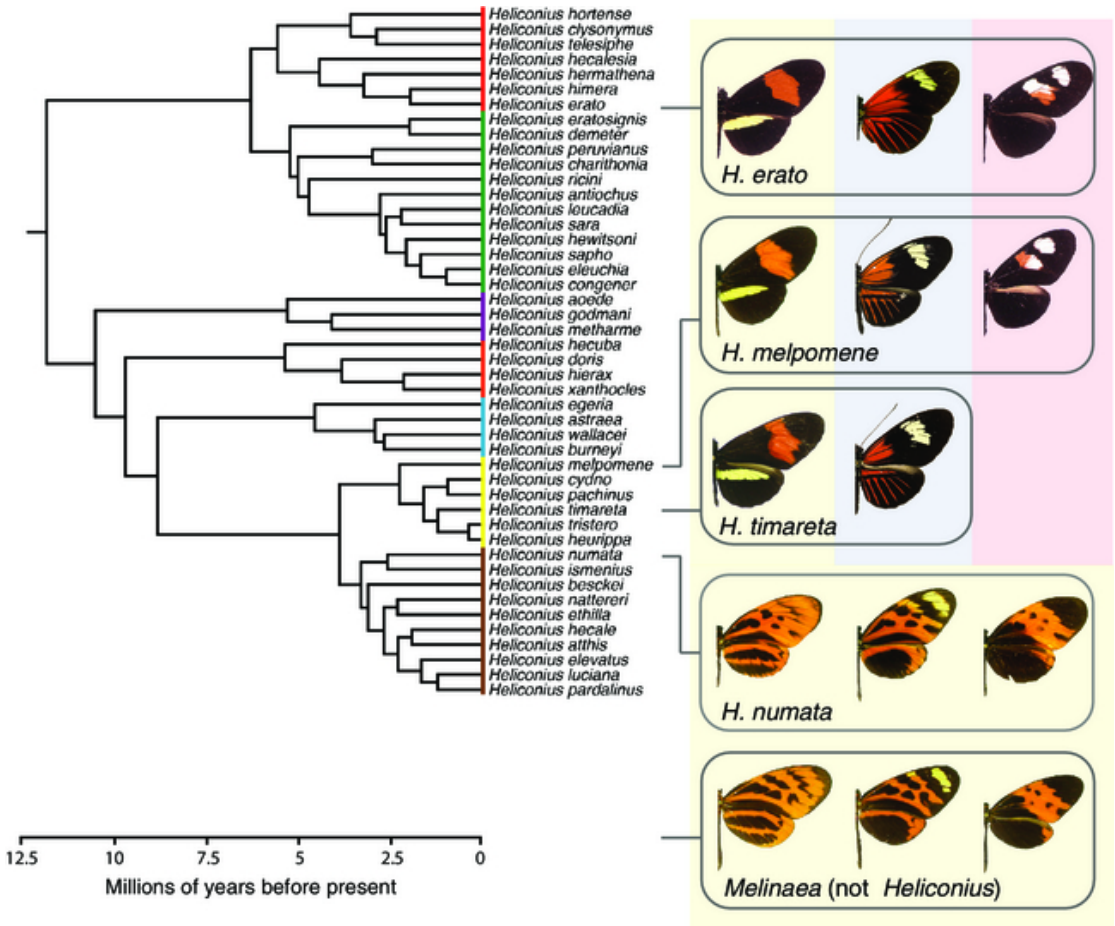


Figure 1. *Heliconius* mimicry in its phylogenetic context

Mimicry is observed between closely related *Heliconius* species (e.g. *H. melpomene* and *H. timareta*), between distantly related *Heliconius* species (e.g. *H. melpomene* and *H. erato*) and between *Heliconius* and heterogeneric species (e.g. *H. numata* and *Melinaea ssp.*). Coloured background boxes indicate taxa that co-occur geographically. Vertical colours indicate subclades: *H. erato* = red; *H. sara and H. sapho* = green; *H. aoede* = purple; *H. doris* = orange; *H. wallacei* = blue; *H. melpomene* = yellow; silvaniform = brown. Phylogeny after *Kozak et al.* (2015). Figure and legend as in Merrill *et al.* (2015).

Heliconius butterflies have been studied for over 150 years and recently, genomic and developmental biology studies have had an important role in the evolutionary debates on the genomic architecture of adaptation and speciation. Major discoveries in evolutionary biology have been possible through the study of *Heliconius* butterflies. Explicitly, by studying *Heliconius* it was possible to gather experimental evidence: 1) for local adaptation maintained by strong natural selection (Mallet and Barton 1989); 2) for widespread gene flow across species barriers (Martin *et al.*, 2013); 3) for horizontal transfer of colour-pattern alleles permitting adaptive introgression (Pardo-Diaz *et al.*, 2012) and evidence that this could lead to hybrid trait speciation (Mavarez *et al.*, 2006; Jiggins *et al.*, 2008); 4) for chromosomal inversions being associated with the evolution of supergenes (Joron *et al.*, 2011); and 5) for divergent warning patterns contributing to assortative mating facilitating speciation with gene flow (Merrill *et al.* 2014); among many others (Merrill *et al.*, 2015). Whether it is true that no single species or clade can be a model for understanding evolutionary trajectories across life, *Heliconius* has undoubtedly contributed to increase our understanding of many evolutionary processes.

Haldane's rule and intrinsic postzygotic barriers

Reproductive isolation is defined as the absence or restriction of gene flow between populations beyond what is caused by spatial separation and can be categorised into prezygotic, extrinsic postzygotic, and intrinsic postzygotic isolation. Intrinsic postzygotic isolation results from genetic incompatibilities and is independent of the environment (Seehausen *et al*., 2014). Two general patterns characterize the evolution of intrinsic postzygotic isolations: 1) hybrid sterility and inviability evolve giving rise to a rough "speciation clock" where there is a either a constant loss in the rate of log hybrid fitness – "linear effect", a de-celerating decrease in the rate of log hybrid fitness – "slowdown effect", or an accelerating decline of the rate of log in overall reproductive compatibility – "snowball effect" (Orr, 1995; Orr and Turelli, 2001; Gourbière and Mallet, 2010); and 2) post-zygotic isolation tends to follow Haldane's rule which states that when one hybrid sex is sterile or inviable that sex is the heterogametic (Haldane, 1922).

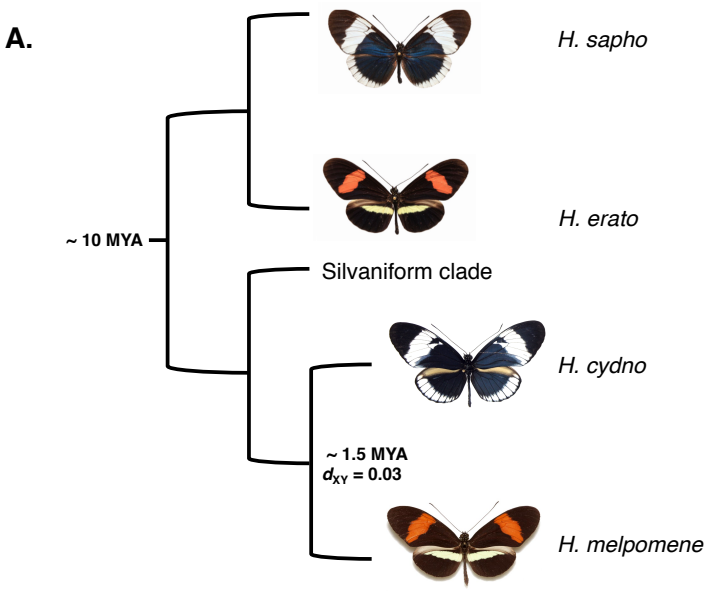
From the first pattern follows that hybrid incompatibilities likely accumulate as a side effect of adaptive or neutral divergence. Genetic theories of speciation have traditionally focused in two explanatory hypothesis for this: 1) the Bateson-Dobzhansky-Muller (BDM) model, which describes postzygotic reproductive isolation as a result of negative epistatic interactions between alleles that have accumulated substitutions while in different genetic backgrounds after being brought into secondary contact (Dobzhansky, 1936; Muller 1940, 1942; Bateson, 1909); 2) the chromosomal model, which invokes the accumulation of rearrangements that result in mis-segregation hybrid backgrounds (White, 1978).

The discovery of hybrid incompatibility genes in different species has led to strong support of the BDM model. For example, a gene encoding a nuclear pore protein, causes epistatic inviability between two fruit fly species, *Drosophila melanogaster* and *D. simulans*. This gene, Nup96, and its protein interactors evolved by positive natural selection in both the *D. melanogaster*

and *D. simulans* lineages (Presgraves et al., 2003). Between these species there is selection driven coevolution among protein interactors leading to incompatible interactions in hybrids. Hence, lethal hybrid incompatibility has evolved as a consequence of adaptive protein evolution (Presgraves and Stephan 2007). However, the evolution of incompatibilities in hybrids does not necessarily need to be a consequence of adaptation or drift in protein coding regions of the genome (Lafon-Placette and Köhler, 2015). For the second pattern, many of Haldane's original examples were actually from Lepidoptera (Haldane, 1922), and Haldane's rule holds across a large range of animals (Schilthuizen et al., 2011).

Haldane's rule has moulded our understanding of the speciation process by laying out a general pattern from which the mechanisms of population divergence and evolution of reproductive isolation can be studied (Orr, 1993; Biddle *et al.*, 1994; Davies, 1996; Brothers and Delph, 2010; Schilthuizen *et al.*, 2011). In *Drosophila*, where most of the research in post-zygotic isolation barriers has historically focused, nearly all cases of inter-species cases of hybrid sterility or inviability are restricted to males (Coyne and Orr, 1998). The pattern holds across the majority of male and female heterogametic taxa and so, Haldane's rule, cannot be explained by the *sensitivity* of one sex over the other and appears to suggest that the genetic mechanisms underlying intrinsic postzygotic isolation could be similar for different systems (Schilthuizen *et al.*, 2011). A composite theory comprising dominance and faster-sex evolution has been proposed as an explanation. First, according to the dominance theory, the majority of hybrid incompatibilities will be partially recessive in hybrids and so recessive sex-chromosome linked incompatibilities will predominantly reduce fitness in the heterogametic sex (Turelli and Begun, 1997; Turelli and Orr, 2000). Second, according to faster-sex evolution, sex-related genes of both sexes evolve rapidly (with a possible a bias towards faster divergence of male-specific factors) (Presgraves, 2002). Together, these two observations, explain both Haldane's rule and the faster evolution of hybrid sterility versus inviability for both male and female heterogametic taxa.

In *Drosophila* and most other species with heteromorphic sex chromosomes, males are the heterogametic sex (XY) and females are the homogametic sex (XX). However, in birds and Lepidoptera females are the heterogametic sex (ZW), and males the homogametic sex (ZZ). Using hybrid cross data from 182 species of Lepidoptera including *Heliconius*, Presgraves (2002) concluded that, as it is observed in *Drosophila*, isolation in Lepidoptera accumulates as species diverge, and sterility precedes inviability. *H. cydno* and *H. melpomene* had their most recent common ancestor 1.5 million years ago and their absolute divergence is approximately 3% (dxy ~ 0.03) (Figure 1A) (Kozak *et al.*, 2015; Martin *et al.*, 2016). Crosses involving several different *Heliconius* species have shown that there is a genetic basis for sterility and that, for example, *H. cydno* x *H. melpomene* F1 female hybrids are always sterile but males are fertile (Naisbit *et al.*, 2002) (Figure 1B). The observed F1 female sterility is concurrent with both the "speciation clock" and Haldane's rule. However, despite strong pre- and post-zygotic isolation barriers there is genome wide evidence of admixture between *H. cydno* and *H. melpomene* (Martin *et al.*, 2013).



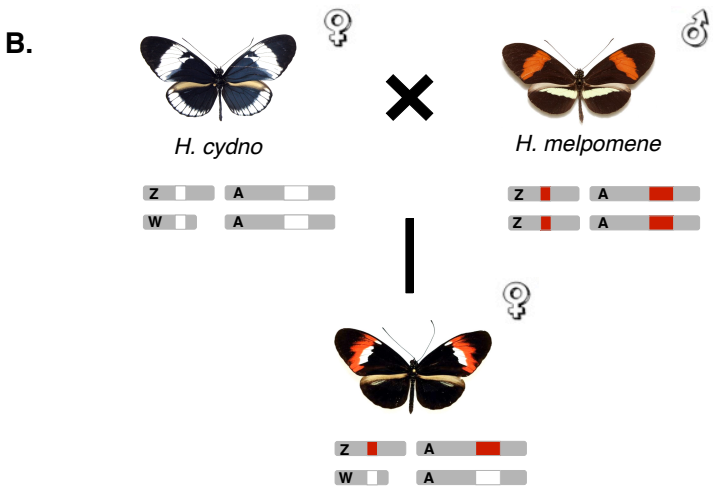


Figure 1. ***H. cydno*** **and** ***H. melpomene***

A. Phylogenetic cartoon of *H. cydno*, *H. melpomene* and outgroups drawn using information from Kozak *et al.* (2015) and Martin *et al.* (2016). Speciation in *H. cydno* and *H. melpomene* is associated with a mimicry shift. There are strong pre-zygotic isolation barriers between *H. cydno* and *H. melpomene* which differ in host-plant, habitat and mate preferences (Jiggins *et al.*, 2001; 2008; Merrill *et al.*, 2011). **B.** *H. cydno* x *H. melpomene* cross cartoon with schematic karyotype for both parental species and female progeny. Post-zygotic isolation barriers are also in place between *H. cydno* and *H. melpomene*. There is strong disruptive selection against the F1 the hybrids which suffer increased levels of predation (Merrill *et al.*, 2012). Moreover, following Haldane's rule, F1 females are always sterile (Naisbit *et al.*, 2002).

The fact that Lepidoptera females are the heterogametic sex and that, unlike *Drosophila* where 20-40% of the genome is X-linked, only 3-5% of the genome is Z-linked (5% in *Heliconius*), offers the opportunity to investigate the evolution of post-zygotic isolation in a different population genetic background (Presgraves, 2002). *Heliconius*, with many species pairs at different levels of

divergence and large quantities of whole genome re-sequencing and transcriptome data available, is an excellent system to investigate the genetic causes of post-zygotic isolation barriers.

Sex chromosomes have unique properties

There are different mechanisms to determine sex, and there is evidence for several evolutionary transitions between the different sex determination systems. Genetic sex determination – widespread in mammals, birds, beetles and butterflies – is likely to have evolved several times from environmental sex determination – widespread in lower vertebrates as fish and reptiles (Charlesworth, 2002). For the transition to occur from environmental to genetic sex determination, genes that override environmental cues needed to be fixed. Close linkage between male and female determining loci is favoured by selection representing the first step in the evolution of highly differentiated sex chromosomes (Úbeda *et al.*, 2015). The suppression of crossing over between proto-sex chromosomes catalysed the reduction of the effective population size of the primitive W/Y, which reduced the strength of natural selection. This results in a reduced ability of natural selection to maintain gene function in the proto-W/Y which gradually degenerated (Charlesworth, 1996). Newly evolving Y or neo-Y chromosomes experience a sharp reduction in effective population size indicating that degeneration can occur over a few million generations (Abbott *et al.*, 2017).

Patterns of variation in DNA can depend on the population genetic properties of the chromosomes on which they reside. Sex chromosomes have unique recombinational and mutational features which alter their evolutionary trajectory. There are several asymmetries between Z and W evolution with respect to the population genetic environment they are in. Recombination is suppressed in the W with the exception of a small pseudoautosomal region and so the W is vulnerable to selective sweeps and selective interference between simultaneously segregating mutations. In a population with an equal

sex ratio there are three times more Z than W chromosomes and so the W is also more susceptible to random genetic drift. Finally, because of the haploid nature of both sex chromosomes in the heterogametic sex, recessive mutations are expected to be under strong selection and patterns of selection for male and female specific gene functions will differ (Ranz *et al.*, 2003; Singh *et al.*, 2014; Grath and Parsch, 2016) (Figure 3).

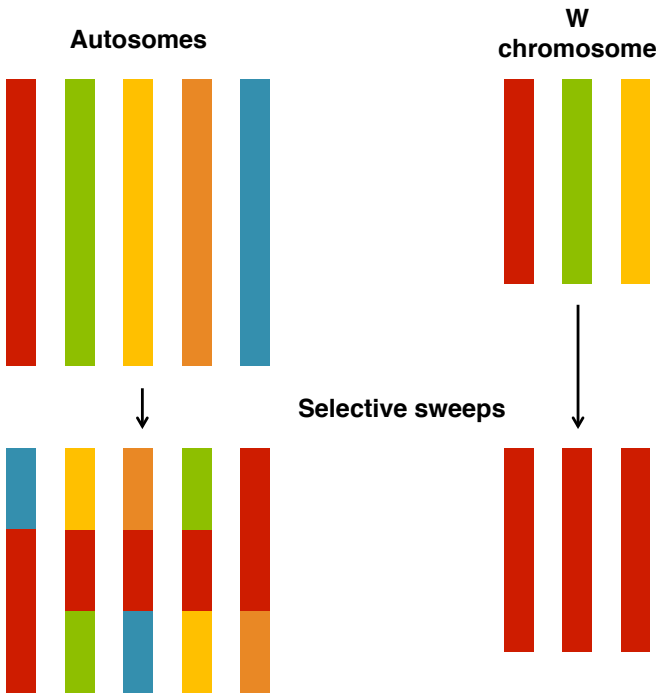


Figure 3. Diversity in autosomes and the W chromosome

Cartoon of autosomes and the W chromosome. Colours symbolise the different alleles present in the population. The greater the number of colours, the greater the degree of genetic diversity. A larger effective population size (N_e) of autosomes compared to sex chromosome allows the autosomes to have a greater degree of genetic diversity. Dominance and recombination also influence the degree of genetic diversity in autosomes and sex chromosomes. The W chromosome is always hemizygous only occurring in the heterogametic sex. In a hemizygous state, recessive adaptive mutations are exposed to more effective selection, which fixes beneficial loci more readily reducing

genetic diversity. The absence of recombination on the W chromosome means that a strong selective sweep would erase diversity more effectively in the W chromosome than in the autosomes. On autosomes recombination around a locus that is subjected to a selective sweep allows diversity to be maintained on other regions of the chromosome. Figure adapted from Ellegren and Galtier (2016).

Sex chromosomes have arisen multiple times and the specific loci involved in sex determination differ dramatically between species. These can involve both coding and non-protein coding regions of the genome. For example, in *Bombyx mori* a single Piwi-interacting RNA regulator is responsible for sex determination (Kiuchi *et al.*, 2014). On the other hand, in *Drosophila*, sex is determined by the activation of the gene *Sex-lethal* (Bell *et al.*, 1988). Differentiation of sex chromosomes results from recombination suppression around the sex determination locus. However, in species like *Heliconius*, in which there is no female recombination, the W chromosome may be non-recombining from the time of origin. Despite several efforts to identify W-linked regions in *Heliconius* involving both whole-genome re-sequencing and transcriptomic data, no W-linked coding or non-coding sequences have been mapped to date.

Further efforts are necessary to identify W-linked regions in *Heliconius*. It is possible that such regions are more difficult to identify in *Heliconius* than in birds due to complete lack of recombination in Lepidoptera females. Moreover, the absence of LINE-1 retro-transposons recognising poly-A tails of mRNAs in avian genomes means that gene transposition events are rare in birds making the W-chromosome have greater synteny to the Z than what is observed for other taxa (International Chicken Genome Sequencing Consortium, 2004). In the W chromosome of the flycatcher, for example, the 46 W-linked genes have paralogues on the Z (Smeds *et al.*, 2015). In *Drosophila*, where males (XY) are also completely non-recombining, even

with a huge investment of time and resources, only fragmentary heterochromatic regions of the *D. melanogaster* Y chromosome have been assembled successfully (Mackay *et al.*, 2012; Carvalho and Clark, 2013).

The evolutionary biology of sex chromosomes addresses many questions including why sex chromosomes exist in the first place; and why are they restricted to multicellular animals and land plants? Moreover, because the W chromosome exists only in females it provides a unique target for the refinement of the functions of genes with female-specific features and so, sex chromosomes, are also relevant in the study of sexual selection, the evolution of sexual dimorphism and evolution of post-zygotic barriers. In addition, sex chromosomes also raise the challenge of understanding how gene expression is regulated to equalise Z-linked expression between the sexes. Finally, the non-recombining sex chromosomes are especially vulnerable to the invasion of TEs promoting sex chromosome distortion in the heterogametic sex (Rinn and Snyder, 2005; Hammer *et al.*, 2008; Mank, 2009; Vicoso *et al.*, 2013).

The homogametic sex carries double the sex-linked genes than the heterogametic and dosage compensation refers to the equalization of gene products between both sexes. Dosage compensation occurs in several animals from mammals (Nguyen and Disteche, 2006), to nematodes (Csankovszki *et al.*, 2004) or insects (Kuroda *et al.*, 2016). However, the mechanisms by which the equalization of expression for sex-linked genes, differ greatly between different organisms to achieve coordinate regulation of the sex chromosomes by differential RNA-polymerase occupancy. In fruitflies, for example, the up-regulation of X-linked genes is mainly male specific (Lucchesi *et al.*, 2005). In birds, a ZW sex determination system, dosage gene compensation can vary for individual genes at different stages of development (Mank and Ellegren, 2009). In *Heliconius* there is one paper published using transcriptomic data to quantify dosage compensation. By quantifying *H. cydno* and *H. melpomene* male and female gene expression the authors concluded that dosage compensation in *Heliconius* is incomplete as there was a slight

increase in male expression relative to female expression (Walters *et al.*, 2015).

Once thought required to balance gene expression levels between sex- and autosomal-linked genes, dosage compensation is far from necessary and it is not required for sex chromosome evolution (Mank *et al.*, 2011; Mank, 2013). Sex chromosome-specific processes such as dosage compensation may affect sex-biased gene expression. However, regardless of dosage compensation, genes that map to sex chromosomes have a different population genetics environment than those mapping to autosomes. Sex-biased genes tend to map disproportionally to the sex chromosomes and may show elevates rates of both protein sequence and gene expression divergence that may in turn drive sexual selection, sexual antagonism or relaxed selective constraint (Grath and Parsch, 2016). In the chapter "Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*", I analyse sex-linked vs. autosomal polymorphism, genetic divergence, and gene expression patterns between males and females, to identify the forces shaping sex and autosome evolution in a female heterogametic system.

Transposable elements shape genomic architecture and can drive reproductive isolation

Nucleotide composition is influenced by biases in mutation and gene conversion, and many other aspects of genomic architecture including the proliferation of mobile genetic elements, arise via drive-like mechanisms (Lynch and Conery, 2003; Galtier *et al.*, 2006; Lynch, 2007). Transposons and retro-transposons (TEs) are mobile genetic elements that have exploited cellular life extremely successfully. The deleterious effects of these genomic parasites cannot be lessened by behavioural avoidance or immunological responses because they reside in the host's genome. Their location guarantees transmission from parent to offspring, and between chromosomal

locations. Mobile genetic elements can also colonize new individuals via horizontal transmission (Sánchez-Gracia *et al*., 2005; Bartolomé *et al*., 2009). Mobile elements are a major source of genomic mutation and their fitness impact on the host is usually negative. However, as with any major source of genomic mutation, they can occasionally increase host fitness (Kidwell and Lisch, 2000).

Since the initial discovery of TEs from the study of unstable mutations in maize many have been described and characterized at the molecular level (McClintock, 1953). The majority of TEs insert at random into the host's genome and so, the most common outcome of a TE insertion, are deleterious mutations. Specifically, TE insertions can create frameshifts and truncations if they insert into coding DNA; changes in gene expression if they insert onto regulatory regions; or can result in large-scale rearrangements. Selection at the level of the TE family favours lineages with greater proliferating ability. However, selection against hosts with high TE loads reduces TE proliferation and drives the evolution of host resistance factors. The level a TE family expands within a host species depends on the relative strength of these two selection pressures and is likely to change through time by a co-evolutionary arms race between the TEs and the host's genome. To sum up, long-term TE success comes from the ability to stabilize a copy number high enough to avoid stochastic loss, but low enough to minimize the risk of host extinction (Brookfield, 1986; Charlesworth, 1987).

There is a large diversity of TEs that differ in their abundance, replication mechanism and in the type of promoter and *cis*-regulatory element they carry. The broadest way to distinguish TEs is based on whether transposition involves an RNA intermediate or not: class I and class II, respectively. TEs are then further subdivided into subclasses, orders and superfamilies. The size of the target site of duplication can be used as a diagnostic feature for most superfamilies (Wicker *et al*., 2007). Transposition can also be classified as autonomous or non-autonomous and there are elements with either type in each one of the two classes. Autonomous TEs encode genes that promote

replication independently of the host chromosomes; non-autonomous TEs require the presence of another TE to be able to transpose. Regardless, TEs still depend on host cell machinery to express their genes and have evolved *cis*-regulatory sequences that mimic the host's promoters (Chuong *et al.*, 2017).

Class I TE transposition involves the production of a processed mRNA transcript that becomes reinserted in the host's genome after being reverse transcribed in complimentary DNA by a reverse transcriptase that is encoded by the element. Therefore, class I TEs are commonly described as transposing through a replicative *copy and paste* mechanism (Wicker *et al.*, 2007). Long terminal repeats elements (LTR elements), long interspersed nuclear elements (LINE) and short interspersed nuclear elements (SINEs) are example of class I transposons with different promoter and *cis*-regulatory elements. LTR elements have RNA polymerase II promoters in *cis* flanking the coding sequence of the element. The mechanism of LTR element replication leads to the introduction of two copies of the LTR in the host. The acquisition of proteins like the ones of endogenous retroviruses by some LTR transposons like gypsy allows the TEs to leave the cell and become infections retroviruses (Malik *et al.*, 2000). On the other hand, LINE have RNA polymerase II at the 5' untranslated region (UTR) and an anti-sense promoter. LINEs usually suffer 5' truncations after insertion removing their promoter sequences. SINE are derived from cellular genes transcribed by RNA polymerase III. SINEs are non-autonomous retrotransposons that are copied by the LINE replication machinery (Chuong *et al.*, 2017).

DNA transposons transpose mostly via *cut and paste* mechanism and do not generate an RNA intermediate during transposition. DNA transposons all belong to class II and have a diversity of element encoded enzymes like transposase or tyrosine recombinase. Transposase makes a cut at the target site producing sticky ends and cutting out the DNA transposon ligating it then to its target site. DNA polymerase fills in the resulting gaps from the sticky

ends followed by DNA ligase. Class II transposons may be identified by short direct repeats followed by inverted repeats. Some DNA transposons transpose not by this *cut and paste* mechanism but by replicative transposition. In these cases a TE replicates itself to a new target site like it is observed for helitron rolling circle (RC) TEs. During the S phase of the cell cycle, if a donor site has already been replicated but a target site has not TEs may become duplicated. Duplications at the target side can result in gene duplication impacting evolution at protein coding regions of the genome (Morgante *et al.*, 2005).

Due to small population sizes, the power of selection in eukaryotes is generally insufficient for the emergence of TE-encoded mechanisms of self-regulation (Lynch 2007). However, the benefits of reducing element activity for the host are much greater than for the TE, and so host mechanisms for regulating TE activity have evolved. Host-encoded mechanisms for reducing TE activity encompass: 1) homology-dependent mechanisms, and 2) transcriptional silencing via methylation (Bourc'his and Bestor, 2004; Slotkin and Martienssen, 2007). One aspect of mobile element evolution that is mutually advantageous to both TE and host is the restriction of element activity to the germline once somatic cell damage will reduce the reliability of the host carriers. This pattern is observed throughout the eukaryotes in *Drosophila* (Bucheton *et al.*, 1976), humans (Gilbert *et al.*, 2002), or *Arabidopsis* (Galli *et al.*, 2003).

While most TE insertions have deleterious fitness impact on their hosts, there are also cases where TE insertions are beneficial to the host. For example, the expansion to near fixation of mutant alleles conferring insecticide resistance in *Drosophila* is a knockout induced by a mobile element insertion (Rostant *et al.*, 2012). TEs are also involved in host chromosome stability in eukaryotes. For example, as *Drosophila* lacks telomerase, a TE became central to telomere stability in this group (Pardue and DeBaryshe, 2011). Mobile genetic element insertions have the potential to modify the activity of

flanking genes. As TEs have their own promoters and regulatory elements, TE insertions into host gene regulatory regions can result in modifications of transcription rates (Wang *et al.*, 2013). Finally, TE insertions can also have structural effects on the products of flanking genes. For example, the extension of retrotransposon transcripts onto downstream genes provides a mechanism for duplicating host-gene sequences, which can eventually evolve novel functions and expression patterns. As such, *Mutator*-like transposable elements (MULEs) in plants, accumulate gene fragments during their transposition and represent and important mechanisms for the genic evolution (Jiang *et al.*, 2004). These examples illustrate that although selection acts to remove TEs from host genomes, there is also the opportunity to for TE family maintenance by positive selection.

All classes and types of eukaryotic TEs have been identified in insects and the repetitive nature of TEs can be used for their discovery (although not all repetitive elements are TEs). In *Heliconius* it has been estimated that TEs comprise roughly 25% of the genome but no study focusing on the TE expression landscape between different species has been carried out (Lavoie *et al.*, 2013). TEs have a profound impact in host's genome organization and stability and TE over-expression can result in sterility. Moreover, TE insertions can also impact expression patterns of flanking genes. Host homology-dependent mechanisms of TE suppression exist in *H. melpomene* and protein coding genes responsible for the activation of such suppression mechanisms have been identified (Lewis *et al.* 2017). In the chapter "piRNA mediated epigenetic silencing does not underlie post-zygotic isolation between *Heliconius cydno* and *Heliconius melpomene*" I focus on whether TE de-repression could explain the sterility phenotype of F1 *H. cydno* x *H. melpomene* female hybrids and whether mechanisms of TE suppression can be responsible for reproductive isolation between *H. cydno* females and *H. melpomene* males.

Gene duplications precede the origin of evolutionary novelty but divergent resolution of duplicate genes can lead to hybrid incompatibilities

Gene and genome duplication are another mechanism by which genomes expand and are thought to play an important role in the evolution of complex phenotypes. The hypothesis that gene duplication is a major mechanism by which evolutionary novelty arises precedes both the molecular and the genomic eras. In 1970, Ohno published *Evolution by gene duplication* where he argued gene duplication is a major mechanism in the origin of novel gene functions (Ohno, 1970). Since, it has been firmly established that duplicate genes are indeed major contributors in the origin of adaptive evolutionary novelties. Many key evolutionary lineages of multicellular eukaryotes have experienced one or more complete rounds of genome duplication and segmental duplications encompassing genes are constant in all organisms (Wolfe, 2001). Examples span all kingdoms of life from bacteria (Blount *et al.*, 2012), to plants (Irish and Litt, 2005), primates (Dulai *et al.*, 1999), fish (Deng *et al.*, 2010) or *Drosophila* (Ding *et al.*, 2010).

There are four mechanisms by which gene duplications arise: 1) unequal crossing-over, where a crossover occurs between two regions with sequence similarity at non-homologous sites resulting in one chromosome with a duplications and another one with a deletion; 2) retrotransposition, where a gene is transcribed along with an upstream retrotransposon is inserted into after reverse transcription of an mRNA intermediate; 3) capture by a double-strand break, where an exogenous fragment containing a genic sequence is inserted into a chromosomal break-point; and 4) ectopic exchange, where a gene copy is generated by strand extension before re-annealing to the broken chromosome (Lynch and Conery, 2000).

The strength of gene duplications as an evolutionary force depends, however, on the rate at which gene duplications arise, and the preservation of duplicate genes by neofunctionalization increases with effective population size. The

most common fate of gene duplicates is pseudogenization through the accumulation of deleterious mutations (Lynch and Force, 2000). To be successful, a duplicate gene must drift towards fixation and selection needs to be sufficiently strong to prevent loss by degenerative mutation. The mechanisms by which duplicate genes are preserved impact genomic evolution at a fundamental level (Lynch 2007). For example, the reciprocal preservation of both members of a duplicate pair leads to an expansion in genome size. On the other hand, the preservation of an unlinked duplicate gene combined with the loss of the ancestral copy does not have an effect on gene number or increases genome size but changes gene order (Lynch and Force, 2000).

Most studies of gene duplications have focused on their potential role in the origin of evolutionary novelty. However, the evolution of a novel function is not the only way a gene duplicate can be preserved. Random genetic drift and degenerative mutations can drive the preservation of duplicate genes by a process known as subfunctionalisation (Oka 1953, 1957, 1974; Lynch and Force, 2000). Through random silencing of paralogues or transposition to unlinked genome position, gene duplication can, therefore, be important in the origin of reproductive isolation (Oka 1953, 1957, 1974; Ting *et al.*, 2004; Bikard *et al.*, 2009). Observed rates of gene duplication indicate that subfunctionalisation can result in near complete genomic incompatibility within a few million years after gene flow stops, which is the approximate time scale over which postzygotic isolation generally occurs in animals (“speciation clock”) (Coyne and Orr, 1997; Presgraves, 2002; Price and Bouvier, 2002). Hence, gene duplications have the potential to drive adaptive phenotypic change and reproductive isolation.

Specifically, duplication events in the sex chromosomes could be relevant to understanding Haldane’s Rule (Orr, 1993). For example, in most mammals autosomal *CDYL* and *CDYL2* have key housekeeping and testes-specific functions. However, in the lineage leading to humans, a copy of *CDYL* was duplicated in the Y chromosome, where it retained the function of

spermatogenesis but lots the housekeeping function, while the autosomal loci loss the function of spermatogenesis but retained the housekeeping role (Dorus *et al.*, 2003). On the other hand, in some *D. melanogaster* x *D. simulans* hybrids, sterility appears to be a simple consequences of the movement of an essential gene to a new chromosomal location via an intermediate phase of gene duplication without any change of function (Masly *et al.*, 2006).

The great majority of research on the genetic mechanisms of species-barriers has focused on *Drosophila* and on the search for "speciation genes" (Coyne and Orr, 1998; Orr *et al.*, 2004; Mallet, 2006). However, adaptive radiations may be associated with duplications events because gene duplications: 1) open up evolutionary pathways for the origin of evolutionary novelties, and 2) generate a population genetics environment that is highly conductive to the passive origin of reproductive barriers. In the chapter "The comparative landscape of duplications in *Heliconius melpomene* and *Heliconius cydno*" I use whole-genome next-generation sequencing data to map duplications among wild-caught *Heliconius* individuals and identify duplicated loci under divergent selection that may play a role in speciation. With the availability of next-generation sequencing data and high quality reference genomes, we can now measure the different evolutionary forces at play after a duplication event in non-inbred lab organisms.

Gene expression in inter-specific incompatibilities

Eukaryotic gene expression is initiated by the recruitment of transcription factors by upstream regulatory elements. These regulatory elements help activate the transcription machinery at the correct initiation site. Transcription must be initiated upstream the translation initiation site and elongation has to ensue far enough to incorporate the translation termination site (Madhani, 2013). A long region between the transcription and translation sites allows to fine tune gene expression at the level of mRNA localization or translation, but

increases mutational rate of origin of premature translation initiation sites and sites for TE insertions (Kim and Jinks-Robertson, 2012). Hence, the benefits of tuning transcription must be weighed against the potential increase of the mutational target size (Lynch and Conery, 2003; Lynch 2007). Although there a lot of research on protein sequence evolution and on the spatial-temporal regulation of gene expression, the evolutionary mechanisms essential to create functional transcripts and how they are shaped by the genomic landscape are poorly understood (Wray *et al.*, 2003).

Inter-specific F1 hybrids have genetic material inherited from both parent species, and hybrid offspring exhibits changes in gene expression that results from the reconciliation of two different genomes and regulatory networks in the same genomic and cellular context (Landry *et al.*, 2005). Quantifying gene expression variability can help to understand 1) gene expression novelty and its possible role in adaptive evolution; as well as 2) the possible link between gene expression differences and maladapted hybrid phenotypes (Wolf *et al.*, 2010). Gene expression changes have been linked to speciation events (Wittkopp *et al.*, 2008) and can drive hybrid speciation when, for example, they enable the colonization of new habitats (Hegarty *et al.*, 2008).

Genetic networks are composed of a large number of interacting genes and are expected to be robust. Robustness of such networks means that both slightly deleterious and advantageous mutations are not expected to necessarily manifest as phenotypic differences (Wagner, 2000). It has been posited that many of the qualitative features of known transcriptional networks could have arisen by non-adaptive evolutionary forces as, amongst others, there is no evidence that genetic pathways emerge *de novo* in response to selective pressures (Lynch, 2007). If divergence in gene expression evolves under neutrality, regulatory hybrid incompatibilities are expected to emerge in a fashion similar to what has been described for the BDM model (Wolf *et al.*, 2010). For example, in *Drosophila*, divergence in gene regulatory regions contributes to the evolution of BDMs (Haerty and Singh, 2006). Hybrid mis-expression is observed in a wide range of taxa and so, this neutral BDM-like

hypothesis, may accurately represent how gene expression evolves (Landry *et al.*, 2005; MAVAREZ *et al.*, 2009; Combes *et al.*, 2015). By identifying such mis-expressed loci it may be possible to map genes correlated to fitness decrease in hybrids.

Studies of hybrid mis-expression have also focused on the relative role of *trans* and *cis* factors in the evolution of novel phenotypes. *cis*-regulatory changes are considered important in adaptive evolution as selection is thought to operate more efficiently on *cis*-regulatory mutations. The reason for this is two-fold: 1) *cis*-regulatory sequences tend to be co-dominant; and 2) modular organization and tissue-specific expression governed by enhancers tends to reduce the level of negative pleiotropy (Wray *et al.*, 2003; Wolf *et al.*, 2010). Examples of *cis*-regulatory mutations with important phenotypic consequences spans sticklebacks (Shapiro *et al.*, 2004), humans (Olds and Sibley, 2003) and maize (Clark *et al.*, 2006). In *Heliconius*, *cis*-regulatory regions modulating the spatial expression of key patterning genes in the developing wing are likely to provide a mechanism for the rapid evolution of novel wing pattern morphologies (Van Belleghem *et al.*, 2017). For the stickleback and *Heliconius* examples these cis-regulating regions are extremely important in adaptation and, consequently, are important in species identity and on the speciation process. In the chapter “Sterility in *Heliconius cydno* x *Heliconius melpomene* F1 female hybrids: a phenotypic and gene expression study of hybrid incompatibilities” I quantify phenotypic and gene expression differences in the parental species and in the hybrids to identify genes that may underlie hybrid sterility. This is the first study of its kind in *Heliconius* and a first step towards the full characterisation of hybrid female incompatibilities.

Conclusion

Speciation is a continuum and quantifying the evolutionary forces driving it is essential to understand the evolution of reproductive isolation. Interactions

between the different evolutionary forces and the population genetic environment in which they exist results in a heterogeneous genomic landscape of divergence. Moreover, inter-specific hybridization during the speciation process also shapes the genomic landscape and further delimits loci with restricted gene flow. Throughout the following chapters I aimed to identify putative barrier loci that reduce gene flow between *H. cydno* and *H. melpomene*. I used genomic approaches to investigate structural variation and transcript variation divergence between the two species. Specifically, I focused on sex chromosome evolution, TE mis-regulation, gene duplication landscape and gene expression differences between *H. cydno* and *H. melpomene*. Many of these analyses had never been done in *Heliconius*, and I hope they make a small contribution to increase our current understanding of the genomic landscape of speciation in *Heliconius.*

CHAPTER 1

The comparative landscape of duplication in *Heliconius melpomene* and *Heliconius cydno*

Abstract

Gene duplications can facilitate adaptation and may lead to interpopulation divergence, causing reproductive isolation. I used whole-genome resequencing data from 34 butterflies to detect duplications in two *Heliconius* species, *Heliconius cydno* and *Heliconius melpomene*. Taking advantage of three distinctive signals of duplication in short-read sequencing data, I identified 744 duplicated loci in *H. cydno* and *H. melpomene* and evaluated the accuracy of the approach using single-molecule sequencing. I found that duplications overlap genes significantly less than expected at random in *H. melpomene*, consistent with the action of background selection against duplicates in functional regions of the genome. Duplicate loci that are highly differentiated between *H. melpomene* and *H. cydno* map to four different chromosomes. Four duplications were identified with a strong signal of divergent selection, including an odorant binding protein and another in close proximity with a known wing colour pattern locus that differs between the two species.

Introduction

Gene duplications occur frequently in eukaryotic genomes, where duplication rates are on the order of 0.01 per gene per million years (Lynch and Conery, 2000). Duplication is considered to be the main mechanism by which new genes arise (Katju, 2012), providing material for the origin of evolutionary novelties (Hunt *et al.*, 1998; Manzanares *et al.*, 2000; Kassahn *et al.*, 2009; Qian and Zhang, 2014). For example, the frequency of gene copy-number variants (CNVs) increased during experimental evolution experiments in *Caenorhabditis elegans* (Farslow *et al.*, 2015) and, in *Escherichia coli*, a tandem gene duplication was responsible for the evolutionary novelty in citrate metabolism seen in the long-term evolution experiment (Blount *et al.*, 2012). Such variation shapes gene expression profiles and influences phenotypic diversity (Feuk *et al.*, 2006; Iskow *et al.*, 2012; Katju and Bergthorsson, 2013).

The most common outcome for gene duplicates is to become pseudogenes through the accumulation of deleterious mutations (Lynch and Conery, 2000). Preservation of duplicate genes by natural selection may depend on whether or not one of the two gene copies accumulates mutations that lead to novel beneficial functions (Ohno, 1970). For example, trichromatic vision in Old World primates evolved by duplication of an X-linked opsin gene, an example of *neofunctionalization* (Hunt *et al.*, 1998). In addition, preservation of gene duplicates by natural selection may also occur by selection for increasing gene dosage as shown for ancient duplicates of *Saccharomyces cerevisiae* (Conant and Wolfe, 2008) or for regulatory robustness (Keane *et al.*, 2014). The duplication event does not, however, need to span the complete length of the gene. For example, a partial gene duplication is responsible for the origin of the antifreeze glycoprotein in Antarctic fish (Deng *et al.*, 2010).

Alternatively, in *subfunctionalisation* models, duplicates are preserved through each copy adopting a subset of the functions of the ancestral gene (Lynch and Force, 2000). This might occur when, for example, regulatory elements of the duplicate loci accumulate mutations that enable both duplicates to take on

new functions different to that of the ancestral gene. In zebrafish, *engrailed-1* and *-1b* are a duplicate pair of transcription factors that evolved complementary expression patterns (Force *et al.*, 1999).

Gene duplication can also contribute to speciation. Duplicate genes can provide the raw material for populations to evolve divergent strategies and adapt to novel habitats, or may lead to genetic incompatibilities (Ting *et al.*, 2004). As such, diversification in gene function between duplicated genes can potentially contribute to reproductive isolation. In *Arabidopsis thaliana* recessive embryo lethality is explained by the divergent evolution of two paralogues of a duplicate gene important for the catalyses of the biosynthetic pathway producing histidine. The reciprocal gene loss has led to genetic incompatibilities in specific crosses (Bikard *et al.*, 2009).

Historically, CNVs were identified with cytogenetic technologies such as fluorescence *in situ* hybridization and karyotyping. More recently, array-based comparative genomic hybridization and single-nucleotide polymorphism array approaches have been used. However, array experiments have several weaknesses including limited coverage of the genome, hybridization noise and difficulty in detecting novel and rare variants (Zhao *et al.*, 2013). It is now possible to detect CNVs using next-generation sequencing technology that generates millions of randomly sampled short (100–300 bp) reads in a single run. Several methods have been developed to detect CNVs from short-read data: (1) analysis of abnormally mapping read pairs (paired-end (PE)); (2) analysis of the number of reads aligned to regions of the genome, or read depth (RD); (3) analysis of clipped/gapped alignments, or split reads (SRs); and (4) *de novo* assembly of resequenced genomes (Ye *et al.*, 2009; Abyzov *et al.*, 2011; Rausch *et al.*, 2012; Chen *et al.*, 2014). In order to increase the accuracy and confidence of the calls, a common approach is to integrate the different strategies into a pipeline where complementary signals are incorporated (Mills *et al.*, 2011; Teo *et al.*, 2012; Tattini *et al.*, 2015; Lin *et al.*, 2015). CNVs have now been surveyed across the genomes of a range of closely related species or populations such as sticklebacks, pea-aphids, pigs

and fruit-flies (Feulner *et al.*, 2013; Chain *et al.*, 2014; Paudel *et al.*, 2015; Rogers *et al.*, 2015; Duvaux *et al.*, 2015).

Here I investigate duplications in the genomes of two species of neotropical *Heliconius* butterflies. This taxonomic group has been studied for over 150 years since the first evolutionists became fascinated with their striking wing pattern diversity. Since then, *Heliconius* has contributed to answering evolutionary questions covering a broad range of research topics from taxonomy to ecology, behaviour and genetics (Merrill *et al.*, 2015). The best studied species pair are *Heliconius cydno* and *Heliconius melpomene*, two hybridizing sympatric species that differ in their ecology, mimicry patterns and mate preferences. They show low levels of inter-specific hybridization that nonetheless results in genome-wide signatures of admixture (Martin *et al.*, 2013). An outstanding question remains over the number and identity of the genomic regions that contribute to their speciation.

Genetic studies of *Heliconius* butterflies have focussed on loci controlling colour patterns, with many races diverging at these loci alone (Nadeau *et al.*, 2011; Martin *et al.*, 2013). Strong and rapid ecological divergence seems to be a driver of the earliest stages of speciation (McMillan *et al.*, 1997; Jiggins *et al.*, 2001; Muñoz *et al.*, 2010). However, recently, gene duplication in the genus has been linked to the evolution of visual complexity, development and immunity (The Heliconius Genome Consortium, 2012), as well as female oviposition behaviour (Briscoe *et al.*, 2013). Moreover, Nadeau *et al.* (2011) identified multiple CNVs between different *Heliconius* races. These results make *Heliconius* butterflies a promising system for an investigation of evolution by gene duplication for both autosomal and sex-linked genes.

I identify duplications using PE, SR and RD information from whole-genome resequencing short-read data for two *Heliconius* species, *H. cydno* and *H. melpomene*, using a similar strategy to the one used to discover and genotype structural variants in the human 1000 Genomes Project (Mills *et al.*, 2011) and the *Drosophila melanogaster* Genetic Reference Panel (Zichner *et al.*, 2013).

By integrating different variant calling algorithms, and taking advantage of three distinctive next-generation sequencing signals, I map duplications among wild-caught *Heliconius* samples from two different species and three different locations, and identify loci putatively under divergent selection that may play a role in speciation.

Materials and Methods

DNA sequence data retrieval and mapping of short-read data

Illumina (San Diego, CA, USA) paired-end sequencing data for 20 *H. melpomene* and 14 *H. cydno* butterflies (SRA106228, Kronforst *et al.*, 2013; ERP002440, Martin *et al.*, 2013) was downloaded from public repositories using the NCBI SRA toolkit (v2.5.7; National Center for Biotechnology Information, Bethesda, MD, USA). The reads were aligned to the *H. melpomene* genome (v2.0) (Davey *et al.*, 2016) with Stampy (v1.0.23) (Lunter and Goodson, 2011) using default values for all parameters except the substitution rate, which was set to 0.01. Picard (v1.128) (picard.sorceforge.net) was used to convert SAM/BAM files and remove PCR duplicate read pairs. Bcftools (v1.3) (Li *et al.*, 2009) and bedtools (v2.20.1-13-g9249816) (Quinlan and Hall, 2010) were used to process BAM and VCF files (Supplementary Table S1).

Detecting duplications through the analysis of SR, PE and RD information

The structural variant discovery methods DELLY (v0.6.1) (Rausch *et al.*, 2012), CNVnator (v0.3.2) (Abyzov *et al.*, 2011) and Pindel (v0.2.5a7) (Ye *et al.*, 2009) were used to detect candidate duplications in a focal set of 10 *Heliconius melpomene rosina* and 10 *Heliconius cydno galanthus* from

Costa Rica in an effort to eliminate sequencing bias in SV discovery, representing the largest population sample available for each species. I ran DELLY and Pindel on each population and CNVnator on each sample individually. These algorithms analyse different sequence signals to call the putative duplications: DELLY uses SR and PE information, Pindel uses SR information and CNVnator uses RD variation. CNVnator was run with a bin size of 100 bp, as recommended by the authors of the software, and all other parameters were set to default values (Table 1, raw calls). There were many more deletions (> 1000s) than duplications after running the discovery pipeline, and I was interested in investigating how might duplications play a role in the adaptive evolution of *H. cydno* and *H. melpomene*. For these two reasons, I focus on duplications and do not report deletions in the resequenced individuals relative to the reference.

The three methods I used to generate our Discovery Sets (PE, RD and SRs) required mapping to a reference genome. Duplication of loci in the reference genome has been shown to influence the discovery of structural variants and the alignment strategy used is important in detecting duplications in repeated regions (Teo *et al.*, 2012). There were several different alignment strategies I could have chosen to deal with reads mapping to more than one location. It was possible to (1) discard these reads, (2) report all possible positions to which the reads map and (3) choose a position at random out of all equally good matching positions.

Limiting the analysis to uniquely mapped regions of the genome (strategy 1) would be likely to miss duplications, especially considering the high heterozygosity of these samples. Using algorithms that consider all possible mapping locations (strategy 2) has not been tested in samples where the mean RD is lower than 20 × (Teo *et al.*, 2012). All the samples I used to generate our Discovery Sets (merged by species) were sequence to an average of 15 × and hence I chose not to use this strategy. Placing a read at

random when all the possible positions are an equally good match (strategy 3) has been shown to dilute the signal of duplications (Teo *et al.*, 2012). However, because this strategy has been used extensively in previous work and is a conservative strategy, I chose this over the other approaches (Zichner *et al.*, 2013).

Species	Method	Raw calls	Merged by tool	Discovery set	Genotyping set	*Heliconius* set
H. cydno	DELLY (PE & SR)	14 691	5 883	1 920	497	744
	CNVnator (RD)	20 936	6 376			
	Pindel (SR)	1 261 451	15 611			
H. melpomene	DELLY (PE & SR)	21 870	5 097	1 591	463	
	CNVnator (RD)	22 267	10 751			
	Pindel (SR)	896 202	7 889			

Table 1. Duplication discovery and genotyping in *Heliconius cydno* and *Heliconius melpomene*

Duplication discovery sets were generated by merging duplications in *H. cydno* and *H. melpomene* using whole-genome resequencing data from 20 wild Costa-Rican individuals (10 *H. cydno galanthus* and 10 *H. melpomene rosina*) (Discovery Set, merged by species). A further 14 wild individuals from Panama (4 *H. cydno chioneus*, 4 *H. melpomene rosina* and 6 *H. melpomene melpomene*) were used to generate each of the species-specific genotyping sets (Genotyping Set). Both

genotyping sets were merged and any resulting redundant calls filtered. This resulted in 744 duplications segregating in the *Heliconius* set.

Filtering and merging duplication predictions: the discovery sets

To generate a list of non-redundant duplications for each species I combined the predictions generated by the three methods using custom scripts (available from Dryad) (Figure 1A). I calculated confidence intervals around each putative breakpoint according to the resolution defined for each method (DELLY: 50 bp outwards, 100 bp inwards; CNVnator: 1 kb outwards, 400 bp inwards; Pindel: +/–10 bp) (Zichner *et al.*, 2013) (Table 1, merged by tool; Figure 1A). I generated six duplication discovery call sets (one for each combination of three methods and two species) by combining all calls with overlapping confidence intervals at both start and end coordinates into a single event. Predictions made by DELLY had to have at least three read-pairs with a mapping quality higher than 20 supporting the call for each individual sample. I removed 311 duplication calls that were predicted by DELLY in all of the *H. melpomene* samples, and were therefore likely to represent either genome assembly errors or genuine deletions in the reference genome.

Finally, I combined the three putative call sets within each species using the intansv module (v1.9.2) in R (v3.2.1) (https://cran.r-project.org; Yao, 2015). I kept calls that had a reciprocal coordinate overlap of 90% or higher and were predicted by at least two methods. Previous studies had used an overlap of 80% (Zichner *et al.*, 2013). However, because the size and total count of the putative variants did not differ dramatically between cut offs of 80 and 90% in our data set (Supplementary Figures S1-S4), I chose to use 90% as a more conservative overlap parameter. The number of duplications that overlap between both species decreases with the increase of the overlap percentage threshold (Supplementary Table S2). Larger duplications are discovered with a higher degree of accuracy (Zichner *et al.*, 2013). If a duplication is a true

positive it is likely to be discovered by more than one sequencing signal (i.e. read-depth, split reads or paired ends). I required that all the duplications in the discovery set were discovered by at least two tools (i.e. two sequencing signals). So, as the percentage of overlap increases, the size of the duplications also increases – larger duplications are more likely to be discovered by more than one tool and are more likely to have the most accurate breakpoints leading to an increase of duplication size with higher percentage overlap between different methods (Supplementary Figures S1-S4, Supplementary Table S2). This generated two species-specific duplication discovery call sets, one for *H. cydno* and one for *H. melpomene (Table 1*, Discovery Set; Figure 1A, Discovery Sets).

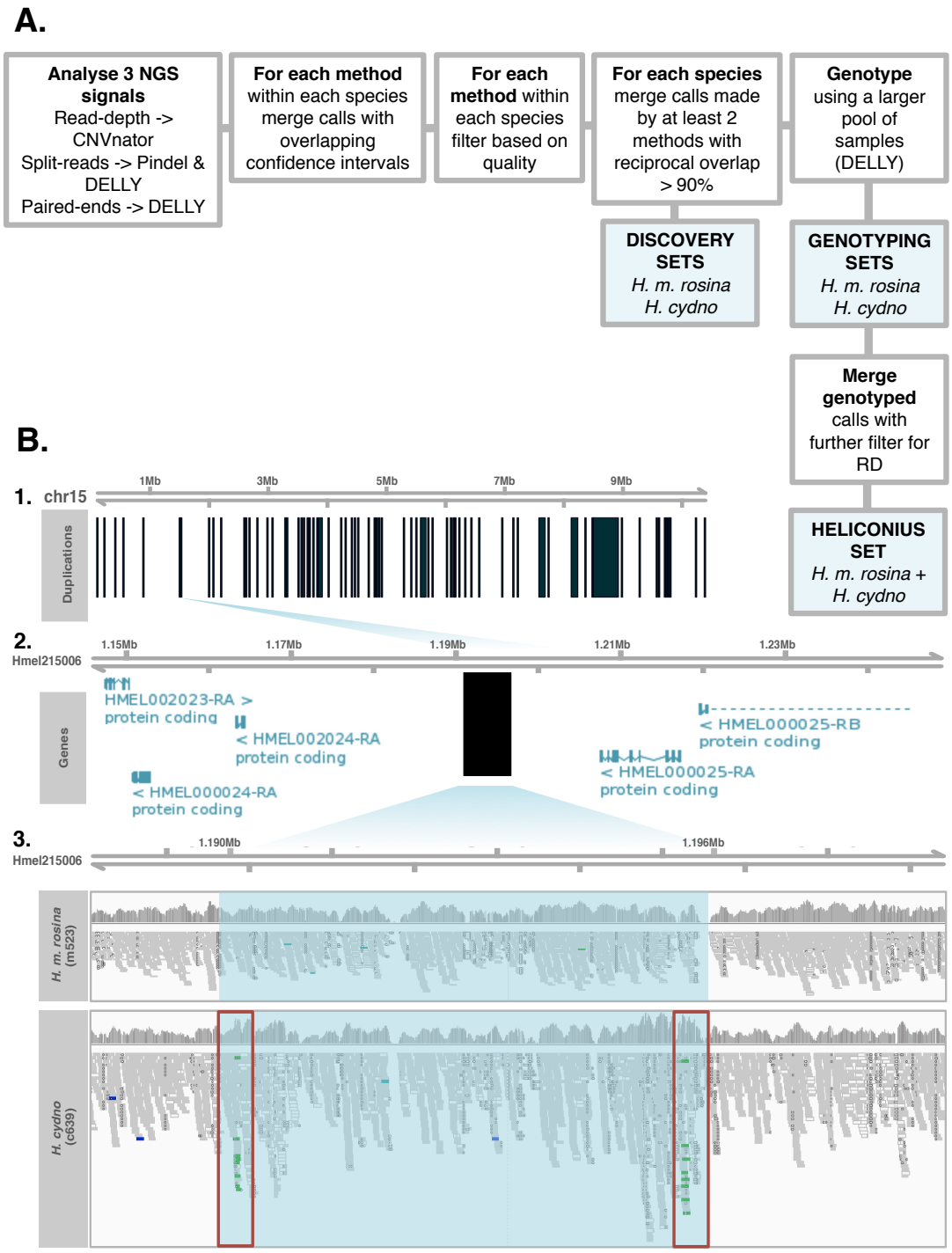


Figure 1. Duplication mapping and genotyping

A. Integrated pipeline for duplication discovery (Discovery Sets) and genotyping (Genotyping Sets). Heliconius Set is the merged and

filtered Genotyping sets from *H. cydno* and *H. melpomene*. **B.** Example of a polymorphic duplication in *H. cydno* with respect to the *H. m. melpomene* reference genome (Davey *et al.*, 2016). **B1.** Schematic representation of merged and genotyped Heliconius set duplication (vertical black rectangles) in Heliconius set for chromosome 15 (Table 1, Heliconius set). **B2.** Zoom-in scaffold Hmel215006 to focus on a putative duplication from the merged genotyped set mapping 5' end of the gene *cortex (Nadeau et al., 2016)* (Table 3, Hmel215006:1190144-1196212). HMEL000025-RA and HMEL000025-RB are transcripts of *cortex* that map to Hmel215006:1205164-1324501. Genes flanking the duplication annotated as in Hmel2 (Davey *et al.*, 2016). **B3.** Zooming-in further and looking at IGV RD and Illumina tracks for one *H. melpomene* and one *H. cydno* sample. Shaded light-blue region delineates the region that was identified as being duplicated. Red rectangles correspond to the breakpoint location of the region. Tracks are coloured green when a tandem duplication with respect to the reference genome is predicted by the read-pair orientation (PE) information. Region displays randomly sampled alignments and track for read coverage depth is not normalised (i.e. not proportional).

Duplication genotype calling: the genotyping sets

To infer copy-number genotypes and evaluate the occurrence of each duplication in both Discovery Sets for all samples (20 *H. melpomene* and 14 *H. cydno*), I used the DELLY genotyper module with –t DUP option and default parameters (v0.7.2) (Rausch *et al.*, 2012). All duplications were treated as dominant loci and genotypes were scored as presence or absence in each sample. Using svprops, a program that computes various SV statistics from an input vcf file (https://github.com/tobiasrausch/svprops), I calculated median read support of each variant. I filtered out duplications with more than 500 reads mapping in an effort to discard repeats found at high copy number

throughout the genome. I also filtered out events not genotyped in any of the samples, leaving high-quality Genotyping Sets of 497 putative duplications in *H. cydno* and 462 in *H. melpomene* (Figure 1A, Genotyping Sets).

Merging the *H. melpomene* and *H. cydno* genotyping sets: the Heliconius set

There were 186 identified putative duplications in the Genotyping Set of *H. melpomene* and *H. cydno* with an overlap >90% and these were merged further using the intansv module (v1.9.2) in R (v3.2.1) (Yao, 2015). After merging both Genotyping Sets according to this criterion I produced the Heliconius Set (Figure 1). Each duplication event was treated as a dominant binary marker (0 for absence and 1 for presence). A duplication was considered to be absent (0) when individual *i* has the same number of copies of sequence *j* as the Hmel2 reference genome, whatever the number of *j* copies in the reference genome. Conversely, a duplication was considered to be present (1) when *i* has more copies of *j* than the Hmel2 reference genome. I called genotypes as presence/absence in this way, rather than calling heterozygotes (Rausch *et al.*, 2012).

Inferring the quality of the putative calls by PacBio alignment and analysis of chromosome 2

I evaluated the accuracy of our duplication calling methods on a separate set of individuals for which appropriate long-read sequence data were available. These were one *H. melpomene* and one *H. cydno* family, for which the parents and one offspring from each family had been sequenced on an Illumina HiSeq 2000 (125 bp paired end, ENA accession ERP009507; see Malinsky *et al.*, 2016 for details). Our full duplication detection pipeline was run on these six individuals for chromosome 2. In addition, pools of 12

female and 12 male larvae from the same two families were sequenced on a Pacific Biosciences (PacBio, Menlo Park, CA, USA) RS II machine (P6/C4 chemistry, ENA submission in progress; read depths: *H. melpomene* females, 54x; *H. melpomene* males, 37x; *H. cydno* females, 49x; *H. cydno* males, 14x).

Pacific Biosciences sequences were aligned to the *H. melpomene* reference genome version 2.0 (Davey *et al.*, 2016) with bwa mem (Li, 2013), using the PacBio option (-x). I then followed Layer *et al.* (2014) to validate our putative duplications, using sambamba (v0.6.1) (Tarasov *et al.*, 2015) to select and filter the SRs from each PacBio bam file and converting these to the bedpe format (v2.25.0) (Quinlan and Hall, 2010) using the LUMPY (https://github.com/arq5x/lumpy-sv) custom script splitReadSamToBedpe. To convert the SRs to breakpoint calls I ran the custom script splitterToBreakpoint on each bedpe file with slope 1000 and default options for all other parameters (Layer *et al.*, 2014). The bedpe files with breakpoint information were merged for each species using bedtools intersectBed (v2.25.0) (Quinlan and Hall, 2010). I selected those reads that overlapped the start and end of the putative breakpoints called using Illumina short-read data. A putative duplication was considered validated when there were split long-read alignments within the predicted breakpoint interval such that (1) two segments of a single PacBio subread aligned to overlapping sections of the reference (Figure 2, PacBio read R1); or (2) if a single read aligned in split formation with the downstream end of the read aligning to a region that is upstream in the reference (Figure 2, PacBio read R2) (Layer *et al.*, 2014; Rogers *et al.*, 2014).

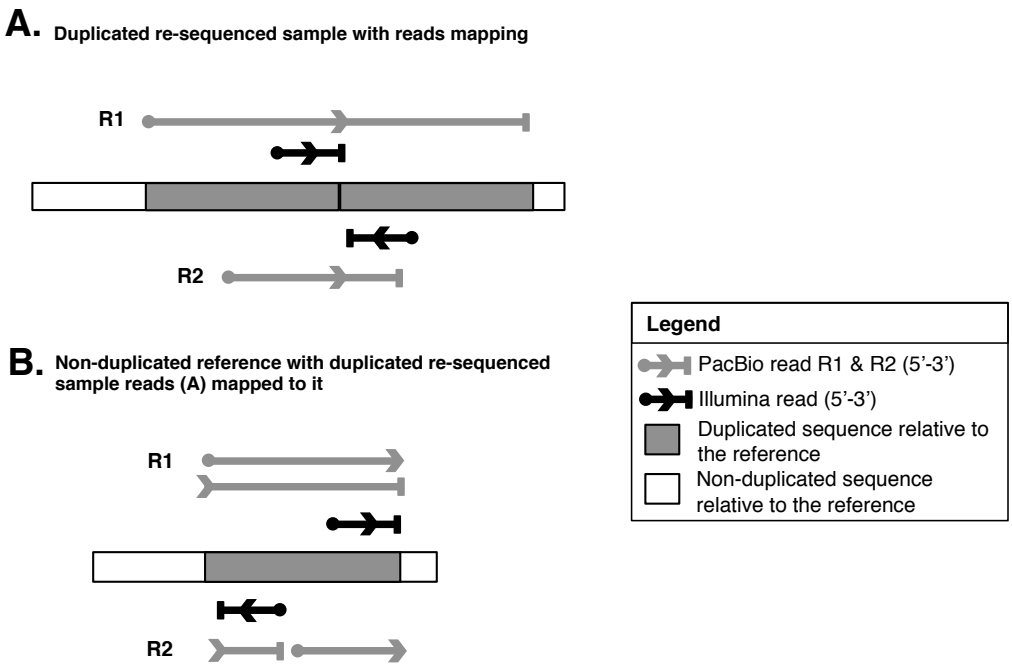


Figure 2. Validating short-read calls on chromosome 2 using PacBio single-molecule sequencing

Example of a breakpoint structure associated with a tandem duplication sequenced by Illumina chemistry (short reads, black) and PacBio chemistry (long reads, grey). A circle denotes the start of a read, the arrow its orientation, and the end is represented by a vertical bar. PacBio read R1 spans the entire duplicated sequence but PacBio read R2 does not. **A.** Duplicated resequenced sample with Illumina and PacBio reads (R1 and R2) mapping. **B.** Non-duplicated reference with duplicated resequenced sample reads from A mapped to it—tandem duplicated sequence aligned to a non-duplicated reference. Illumina reads from an individual with a tandem duplication map in divergent orientations when aligned to a reference without duplicated sequence. When PacBio read R1 is aligned to a non-duplicated reference, there are two alignments to the region that is flanked by the Illumina divergently oriented reads. The PacBio read R2 aligns discontinuously

to the reference genome. The 3′ end of the R2 fragment of the breakpoint aligns to the reference upstream of the 5′ end of the R2 fragment.

Using the putative genotyping duplication call set to show population structure and differentiation

Putative duplications from the Heliconius Set were analysed as dominant loci by principal component analysis in using the R package adegenet (v1.3-1) (Figure 4) (Armengol *et al.*, 2009; Jombart and Ahmed, 2011).

Overlap between structural variants and genomic features

I investigated the overlap between the genotyped duplications and four different genomic features (genes, coding sequences (CDSs), introns and untranslated regions (UTRs)) using the R package ‘intervals' in both Genotyping sets (Figure 1A and Table 1, Genotyping set). A single duplication could fall into several subcategories. All duplications that overlapped with coding sequence were counted as CDS duplications. A duplication was considered to be intronic if it overlapped with an intron but not CDS. UTRs were considered in the same way as introns if it does not overlap with CDS. Overlap with any of these features was considered a gene-overlapping duplication. As a small number of the genotyped duplications were overlapping, these were merged for this analysis, so that only non-overlapping duplication intervals were considered.

To investigate whether the observed number of duplications overlapping each class of genomic features was significantly larger or smaller than expected by chance, I simulated 10 000 randomized distributions of duplications across the genome. In each simulation, the defined set of duplication intervals (with overlapping intervals merged for simplicity) was randomly permuted into non-

overlapping locations across the genome, and the number overlapping with each class of genomic feature was recorded. I used the 2.5 and 97.5% quantiles of the simulated distribution as critical values to assess whether the observed overlaps differed significantly from that expected under a random distribution of duplications.

Detection of enriched biological functions within the Heliconius Set

I used InterProScan (v5.18.57.0; https://www.ebi.ac.uk.uk/interpro/)) (options –t n –goterms) to compare the Heliconius Set against the InterPro database. The InterPro database integrates predictive information from a number of sources (Mitchell *et al.*, 2015). I analysed PANTHER (http://www.pantherdb.org) database IDs that can be used to infer the function of uncharacterized genes based on their evolutionary relationships to genes with known functions (Mi *et al.*, 2016). I ran the PANTHER overrepresentation test on the Heliconius Set using the *D. melanogaster* genome as the reference list. I performed this analysis on the PANTHER GO-Slim Biological Process. I used the Bonferroni correction for multiple testing and report those categories overrepresented with $P<0.05$ (Supplementary Table S3 available online doi:10.1038/hdy.2016.107; and Supplementary Figure S13). Five hundred and twenty nine overrepresented occurrences did not have a biological process associated with them but I have reported their predicted family name (Supplementary Table S4 available online doi:10.1038/hdy.2016.107).

Identifying outlier loci from the Heliconius Set

Duplications present in the Heliconius Set were tested for signals of divergent selection by identifying F_{ST} outliers using BayeScan (v2.1) (Foll and Gaggiotti,

2008) with default parameters except that prior odds were set to 1 (Cheang *et al.*, 2013). F_{ST} was estimated for the Heliconius Set between (1) *H. cydno* Costa (Rica and Panama); and (2) *H. melpomene* (Costa Rica, Panama and French Guiana). Each duplication event was treated as a dominant binary marker (0 for absence and 1 for presence). I corrected for false positives (false discovery rate of $P<0.05$). Duplications with log posterior odds >1 have strong support for selection.

I also applied a related method that identifies loci subject to selection taking into account associated population/species-specific covariates, using BayPass v2.1 (http://www1.montpellier.inra.fr/CBGP/software/baypass/), for the putative duplications in the Heliconius Set (Gautier, 2015). The duplication events were considered as dominant binary markers. I used country coordinates and species as population-specific covariates. The covariates were defined as follows: Costa Rica: 9.7489, 83.7534; Panama: 8.5380, 80.7821; French Guiana: 3.9339, 53.1258; *H.* cydno: 1 and H*. melpomene*: 2. Under the Standard Covariate Model I estimated for each duplication event the Bayes Factor, the empirical Bayesian *P*-value and its underlying regression coefficient using an Importance Sampling algorithm. I simulated the data under the Inference Model to calibrate the neutral distribution of XtX. XtX was used to identify loci subjected to adaptive divergence. After calibrating XtX I ran the Markov chain Monte Carlo algorithm using posterior estimates available from the previous analysis and I corrected for location using just one covariable at a time, as suggested by (Gautier, 2015). Finally, I selected the duplication events that had observed XtX estimates above the 98% threshold of the simulated data (XtX > 7.9). I cross-referenced the regions selected from BayeScan and BayPass analyses to look for overlaps between the two methods.

Results

Duplication maps for *H. cydno* and *H. melpomene*

I identified a Discovery duplication set of 1920 putative *H. cydno* duplications and 1591 putative *H. melpomene* duplications (Table 1, Discovery set: merged by species) based on whole-genome resequencing data from 10 wild *H. cydno* samples and 10 wild *H. melpomene* samples (Kronforst *et al.*, 2013) (Supplementary Table S1). I genotyped the discovery sets in a further 10 *H. melpomene* and 4 *H. cydno* samples (Martin *et al.*, 2013). After removing duplications with low-quality genotypes and high RD and duplications where all samples differed from the *H. melpomene* reference genome, I retained 497 putative *H. cydno* duplications and 463 *H. melpomene* duplications (Table 1, Genotyping set; Figure 3 and Supplementary Figures S5 and S6). I then merged redundant duplications in the *H. cydno* and *H. melpomene* Genotyping Sets, where two variants overlapped in over 90% of their total length, to produce the Heliconius Set containing 744 duplications ranging in size from 228 bp to 207 510 bp (median 5693 bp) (Table 1, Heliconius set; Supplementary Figures S7-S9).

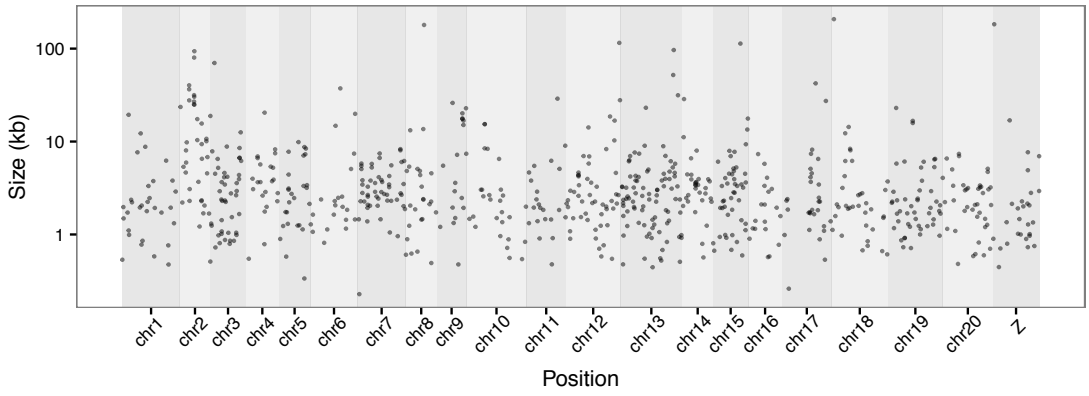


Figure 3. Distribution of the *Heliconius* duplication set mapped to the Hmel2 reference genome

H. cydno and *H. melpomene* genotyping sets were filtered and exclude duplications with a median read count of > 500 reads per sample or not genotyped in any of the samples. The two high-quality genotyping sets

were merged to produce the Heliconius duplication set (Heliconius Set, Figure 1A and Table 1). Each putative duplication on the Heliconius set is represented by a point according to position in the genome (x axis) and size (kb).

Validation rate as estimated by analysis of PacBio single-molecule long reads

I validated our pipeline using Illumina and PacBio sequencing data for a single chromosome from two families of *H. melpomene* and *H. cydno*. I first ran our pipeline on the Illumina data for chromosome 2 and then validated the calls using the PacBio data. Using the Illumina sequenced trio, I identified 97 duplications on chromosome 2 in *H. melpomene* and 137 in *H. cydno* after filtering. I validated 96.9% of the *H. melpomene* and 95.6% of the *H. cydno* calls using single-molecule PacBio SRs for each species separately. I also ran the Heliconius Set of duplications using the same PacBio data, combining the data from *H. cydno* and *H. melpomene*. This confirmed 65.5% of putative duplications.

The lower validation rate on the Heliconius Set duplications is because of the fact that these are different individuals and populations compared with our PacBio data. In the Heliconius set a third to a quarter of all duplications identified only occurred in a single individual and hence were unlikely to be present in the PacBio data (Supplementary Figure S8). Nonetheless, the high validation observed in our reference trios suggests that our pipeline is correctly identifying duplications from Illumina data.

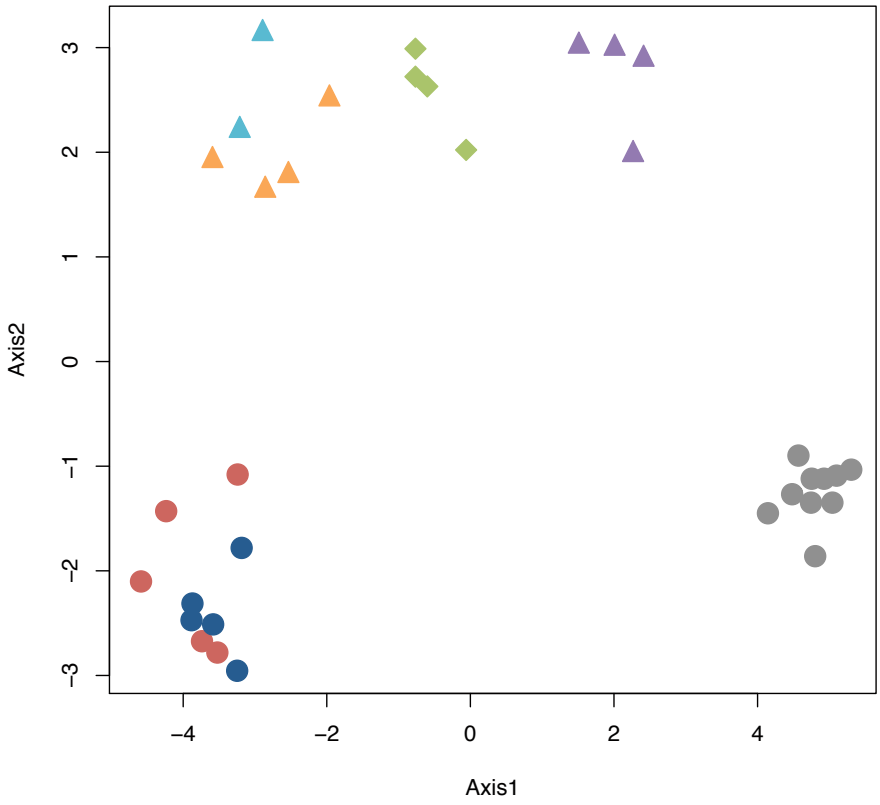
Effect of genome structure on duplication distribution

Most duplications occurred in a small number of samples and there were only a few duplications at high frequency among all the samples (Supplementary

Figure S8). For example, in the *H. cydno* genotyping set, 26.8% of the duplications are singletons and, in the *H. melpomene*, 32.5%. The number of duplications per chromosome in the Heliconius Set is not equally distributed along the different chromosomes (Supplementary Figure S9A) and is weakly correlated with chromosome size (r^2=0.344; Supplementary Figure S9B). There was also variation between individual chromosomes in the number of duplications per Mb ($F(20,723)=14.2$, $P<0.001$). Chromosome 18 tended to have fewer duplications, whereas chromosome 17 showed an excess of duplications per Mb compared with other chromosomes (*post hoc* Tukey's HSD (honest significant difference) test with correction for multiple testing). I did not observe any excess or depletion of duplication events towards the centres of chromosomes in the Heliconius Set (Supplementary Figure S10).

Principal component analysis of the genotyped *H. cydno* and *H. melpomene* sets

I tested for population structure in the Heliconius Set of duplications genotyped as dominant markers using principal component analysis. In total, 17.57% of the total variance was explained by the first two principal components (PCs; PC1 12.97% and PC2 4.6%). Along PC1 the samples separated by species and geography (Figure 4), with all populations distinct except *H. m. melpomene* and *H. m. rosina* samples from Panama that are known to be genetically very similar (Martin *et al.*, 2013). However, PC2 separates the Costa Rica samples from those from Panama and French Guiana. It seems most likely that this is a methodological artefact because samples from different countries came from different sequencing runs (Supplementary Table S1). In addition, our call set was generated from the Costa Rica data set, and subsequently genotyped on both sample sets. Within Costa Rica, PCA analyses separate populations by geography and species as expected (Supplementary Figure S11).



	Species	Location
◆	*H. m. melpomene*	French Guiana
▲	*H. m. melpomene*	Panama
▲	*H. m. rosina*	Panama
▲	*H. cydno chioneus*	Panama
●	*H. cydno galanthus*	Costa Rica
●	*H. m. rosina*	Costa Rica (Pacific)
●	*H. m. rosina*	Costa Rica (Caribbean)

Figure 4. Principal component analysis of the duplicated variants in the *Heliconius* set

Samples cluster by species and location based on their duplication genotype. Of the total variance, 17.57% was explained by the first two principal components (PC1 12.97% and PC2 4.6%).

Overlap between duplication and genes

I found that the genotyped duplications in *H. melpomene* overlapped with genes and CDSs significantly less often than expected by chance, whereas the rate of overlap with UTRs and introns did not differ from the null expectation under a random distribution (Table 2 and Supplementary Figure S12. This is consistent with the idea that duplications involving functional regions have a greater probability of being deleterious, and are therefore more likely to be removed by selection. In contrast to *H. melpomene*, in *H. cydno*, there was no significant deviation from the null expectation in the rate of overlap between genotyped duplications and genes, CDSs, UTRs or introns.

		H. melpomene	*H. cydno*
Complete gene	#	23	41
	%	5.2	8.9
	< Sim 2.5 %	No	No
Gene	#	157	210
	%	35.3	45.8
	< Sim 2.5 %	Yes	No
CDS	#	92	154
	%	20.7	33.6
	< Sim 2.5 %	Yes	No
Intron	#	45	42
	%	10.1	9.2
	< Sim 2.5 %	No	No
UTR	#	27	20
	%	6.1	4.4
	< Sim 2.5 %	No	No

Table 2. Functional impact of the *Heliconius* set

Observed absolute counts and proportion of duplications overlapping complete genes, genes, CDS, introns and UTRs. < Sim 2.5% column indicates whether the observed proportion of overlap with each category falls within the 2.5% confidence interval of the simulated data overlap after 10 000 iterations. If < Sim 2.5% is 'No', then duplication counts are not within the 2.5% confidence interval and the overlaps observed do not significantly differ from random expectations. If 'Yes', then counts are within the 2.5% confidence interval and the overlap observed is significantly less than expected under a random distribution. A single duplication can fall into several subcategories. Abbreviations: CDS, coding sequence; UTR, untranslated region.

Enrichment of biological functions in the Heliconius Set

The duplications I have identified are not equally distributed across the genome (Figure 3, Supplementary Figure S9). The heterogeneity observed across the landscape is likely to be a reflection of biases in the rates at which duplications arise in certain regions or a bias in the preservation of duplications in specific functional classes because of the action of natural selection. It has been shown that multigene families, specifically those involved in environmental responses, are particularly prone to being duplicated/retained (Duvaux *et al.*, 2015). I detected 19 gustatory receptors that had been previously identified as putatively duplicated by CNVnator analysis (Briscoe *et al.*, 2013). Moreover, I tested whether any biological functions were overrepresented in the Heliconius set of duplications using PANTHER (Supplementary Figure S13).

Within the *Heliconius* set there were 1710 different family classes of which 1181 were associated with predicted biological processes. Of these processes, 26 different biological function categories were identified as overrepresented in the *Heliconius* set based on the *D. melanogaster* reference list (P<0.005) (Supplementary Figure S13 and

Supplementary Table S3 available online doi:10.1038/hdy.2016.107). These were involved in transketolase, phosphatase, endodeoxyribonuclease, metallopeptidase, lipid transport, deacetylase, oxidoreductase and transferase activity. There was also a set of 529 family classes that are overrepresented in the Heliconius set but do not have a specific Gene Ontology (GO) term, biological or specific molecular function associated with them but include ejaculatory bulb-specific protein, male sterility protein, cuticle formation and transposable element related (Supplementary Figure S13, Unclassified; Supplementary Table S4 available online doi:10.1038/hdy.2016.107). Structural constituents of the cytoskeleton, protein binding, DNA binding transcription factor and kinase activity were molecular function categories underrepresented in the *Heliconius* set. The biological function that was most overrepresented in the entire set was the GO category related to the pentose-phosphate shunt (primary metabolic process, fold enrichment 18.35, P=5.4e–07). Immune system processes were underrepresented in our set (fold enrichment <0.2, P=2.59e–04).

Identification of outlier duplications in the Heliconius Set potentially under selection

To characterize patterns of divergence observed between *H. melpomene* and *H. cydno* I first calculated F_{ST} between the two species and identified candidate outlier regions using BayeScan for the Heliconius Set of duplications, treating putative duplications as co-dominant (presence/absence) markers. After correcting for false positives I found nine duplications that are candidates for selection (Supplementary Figure S14A and Supplementary Table S5). I also ran BayPass that conducts a similar test by accounting for sample location and species. This produced six putative duplicated regions above the simulated significance threshold (Supplementary Figure S14B and Supplementary Table S5), four of which were also identified by BayeScan (Table 3). I consider the four outlier events found by both tests

to be strong candidates for directional selection. One region, on chromosome 15, is located in an intergenic region upstream of the gene *cortex* that is involved in the regulation of yellow and white wing pattern elements (Figure 1B) (Nadeau *et al*., 2016). The other three regions overlap with genes, predicted to be a Kazal-type serine protease (chromosome 9), an odorant binding protein (chromosome 18) and a regulator of the cell cycle and nitrogen compound metabolic processes (chromosome 21) (Table 3). All four candidate selected duplications are absent in the *H. melpomene* samples and present in 13 or 14 of the 14 *H. cydno* samples.

Chr	Scaffold	Start	End	Size	BayeScan log10(PO)	BayPass mean XtX	Freq in *H. melpomene*	Freq in *H. cydno*	PANTHER GO-Slim Biological process	Hmel2 annotation
9	Hmel209007	4 344 840	4 364 959	20 119	1.7222	7.95239143	0	0.93	Kazal-type serine protease inhibitor	HMEL009267
15	Hmel215006	1 190 144	1 196 212	6068	1.8414	8.78515118	0	1	NA	upstream of *cortex*
18	Hmel218003	221 730	429239	207 509	1.894	8.75630075	0	1	Protein targeting Intracellular protein transport Transport Localization Biological regulation Asymmetric protein localization	OBP41 HMEL013558 HMEL013559 HMEL003174 HMEL003175 HMEL003862 HMEL003863
21	Hmel221012	779 541	796 444	16 903	1.72	8.35788884	0	0.93	Regulation of the cell cycle Regulation of biological process Porphyrin-containing compound Metabolic process Nitrogen compound metabolic process Regulation of translation Primary metabolic process mRNA transcription Nucleobase-containing compound metabolic process Cell differentiation, developmental process Regulation of transcription from RNA pol II promoter	HMEL016617 HMEL016621 HMEL016620

Table 3. The four duplications in the Heliconius set identified as outliers by BayeScan and BayPass analysis

Chromosome position, scaffold name, start, end and size of each putative duplication are indicated. log10 (Posterior Probabilities) from the BayeScan analysis is indicated per duplication between the *H. melpomene* and *H.* cydno. All these loci had positive values of α that suggests diversifying selection. BayPass XtX mean for each loci is also indicated for each species after correcting for location. Allele frequencies calculated as co-dominant markers are shown for each species at the loci (genotyped by Delly2). PANTHER GO-Slim biological processes and Hmel2 annotations retrieved from Hmel2.gff (Davey *et al*., 2016). Abbreviation: NA, not available.

Discussion

Gene duplication is an important source of genetic fuel for evolutionary diversification, and can also contribute to speciation. Here I have used short-read genome sequence data to identify signatures of CNV in natural populations. I have used single-molecule sequencing to validate our pipeline, with a validation rate of ~96% within families. I have successfully identified 744 loci and genotyped them (presence/absence) in 34 wild individuals sampled from the two species *H. melpomene* and *H. cydno*.

Despite the ubiquitous nature of duplications, different chromosomes might be expected to contribute differently to the overall duplication landscape. Large chromosomes tend to have the highest absolute duplication counts but chromosome size is not the sole predictor of duplication distributions. Sex chromosomes, which have more repetitive content, smaller population sizes and lower levels of background selection than autosomes, have been shown to have a higher duplication load per base pair than autosomes in *D. simulans* and in *D. melanogaster* (Mackay, 2010; Charlesworth, 2012; Zichner

et al., 2013; Rogers *et al.*, 2014; Rogers, 2015). However, the X chromosome of *Drosophila yakuba* does not contain an excess of duplications compared with the autosomes and no signals of adaptation through duplication have been identified. Similarly, the *Heliconius* duplication set does not harbour an excess of duplications on the Z chromosome compared with the autosomes. It is possible that duplications are more difficult to detect on the Z chromosome that has higher divergence than the rest of the genome (Martin *et al.*, 2013) and higher proportion of repetitive content (Conrad and Hurles, 2007). Further work will be needed to compare the landscape of duplications across sex chromosomes.

Duplications are not homogenously distributed across the genome (Figure 2 and Supplementary Figures S5 and S6). There was no bias towards telomeric regions as it has been documented for humans (Zhang *et al.*, 2005). *Heliconius*, like *C. elegans*, have holocentric chromosomes and, to our knowledge the enrichment of structural variations in telomeric regions (and/or pericentrimeric regions) has yet to be documented for organisms with this chromosomal organization (Farslow *et al.*, 2015). The number of singletons identified in our data set (a quarter to a third of all duplications) is on the same order of magnitude as that seen previously. For example, Duvaux *et al.* (2015) reported 31% singletons in pea-aphid clones.

A large proportion of structural variants arising in genomes are slightly or moderately deleterious and therefore experience purifying selection (Emerson *et al.*, 2008; Zichner *et al.*, 2013). In *D. melanogaster*, fewer duplications were found in coding sequence as compared with random expectation (Zichner *et al.*, 2013). Consistent with this, I found that in the *H. melpomene* Genotyping Set duplications are biased away from coding regions, although they are not biased away from or towards intronic or UTR regions. However, I did not find a similar bias in *H. cydno*, and saw no significant depletion of the number of duplications in *H. cydno* as compared with *H. melpomene*. This goes against expectations, given that the effective population size of H. *cydno* has been inferred to be around four times greater than that of *H. melpomene* (Kronforst

et al., 2013), consistent with the significantly higher genome-wide heterozygosity in *H. cydno* (Martin *et al.*, 2013). Therefore, I might expect selection to operate more effectively and duplications to be more efficiently removed from *H. cydno*, but this does not appear to be the case. I do not have any good explanation for this.

Although most structural variants may be deleterious, there is particular interest in those few that have positive effects. There are now many examples in which gene duplicates provide the genetic fuel for adaptation, and have been shown to be under positive selection (Beisswanger and Stephan, 2008; Arroyo *et al.*, 2012; Blount *et al.*, 2012). Here, I am specifically interested in speciation. Gene duplicates have been implicated in reproductive isolation for both animals and plants. For example, the Odysseus gene that causes hybrid sterility between *D. mauritiana* and *D. simulans* is a duplicate of the *unc-4* gene (Ting *et al.*, 2004). In *A. thaliana*, paralogues of an essential duplicate gene that evolved divergently interact epistatically in some interspecific crosses and control a recessive embryo lethality (Bikard *et al.*, 2009). In the context of *Heliconius*, I am specifically interested in speciation and divergent selection between the closely related species, *H. melpomene* and *H. cydno*. Using BayeScan and BayPass I identified a relatively small number of duplications that are putatively divergently selected between these species.

Many functionally important regions in different genomes have been documented to evolve through gene duplication followed by neo or subfunctionalization. Genes responsible for environmental response are known to be overrepresented as duplicated sequences in a range of organisms from humans to fruit flies and butterflies (Johnson *et al.*, 2001; Tuzun *et al.*, 2005; Hahn *et al.*, 2007; Briscoe *et al.*, 2013) and in line with previous studies I have detected an enrichment of genes involved in sensory perception (Briscoe *et al.*, 2013; Rogers *et al.*, 2014; Paudel *et al.*, 2015; Duvaux *et al.*, 2015). For example, I detected gustatory receptors that had already been identified in *Heliconius* (Briscoe *et al.*, 2013) but I also detected others such as olfactory receptors and olfactomedin-related proteins

(Supplementary Table 3). Specifically, in our outlier analysis there is an odorant binding protein that is divergent in copy number between *H. cydno* and *H. melpomene* (OBP41, Table 3). Several hypotheses have been put forward to explain the trend of increased CNV among genes involved in environmental response. On one hand, these CNVs might be maintained by positive selection as outlier analysis-based methods have shown an enrichment for these GO classes (Paudel *et al.*, 2015; Rogers *et al.*, 2015; Duvaux *et al.*, 2015). On the other hand, these differences could occur simply because certain sequence motifs like non-B DNA forming sequence are more common in gene-rich regions and, at the same time, they increase the rate of CNV formation (Sjödin and Jakobsson, 2012). Gene categories overrepresented in CNV are also enriched within segmental duplications, and segmental duplications are very structurally dynamic (Conrad and Hurles, 2007). Moreover, families with multiple paralogues are more prone to further copy number variation (Hastings *et al.*, 2009).

Not all the putative duplications I found as outliers were involved in environmental response. Another candidate locus under divergent selection was found near the *cortex* gene that controls the yellow hindwing bar and white/yellow forewing patterns that differ between *H. m. rosina* and *H. cydno* (Nadeau *et al.*, 2016). Moreover, I have also found an enrichment of male reproductive proteins in the Heliconius Set (Supplementary Table S4 available online doi:10.1038/hdy.2016.107). These proteins evolve rapidly and are commonly duplicated in, for example, *D. yakuba* (Rogers *et al.*, 2014). It was somewhat surprising, however, that I did not observe an enrichment for immunity-related genes.

Interestingly, the four putative duplicated regions I have identified as excessively differentiated in *H. cydno* and *H. melpomene* were all nearly fixed in *H. cydno* but not in *H. melpomene. H. melpomene* and *H. cydno* differ in many aspects of their ecology and behaviour. Shifts in host plant have played a central role in their diversification. The evolution of host-use strategies reflects a trade-off between selection pressures (Merrill *et al.*, 2013). For

example, gene duplications that persist in an evolving lineage have often been found to be beneficial because of a protein dosage effect in response to environmental conditions. Host-plant systems may be subject to rapid coevolution and duplicated loci in *H. cydno* could be related to the fact that *H. cydno* is a host plant generalist and *H. melpomene* is a specialist (Merrill *et al.*, 2013). The Fst values may to be underestimated in our analyses. It is possible that some of the variants identified have greater *true* Fst values than reported in this study and, due to lack of power because the markers were treated as dominant (Lynch and Milligan 1994), did not pass the significant threshold required (Table 3).

The duplications I have identified as being under selection between *H. cydno* and *H. melpomene* may play a role in species divergence. I have shown that, despite being ubiquitous, the landscape of duplications in *Heliconius* is heterogeneous and likely to be under both positive and negative selection. The putative duplications I found merit further investigation for their potential role in host plant and mate recognition differences between the species.

Supplementary Tables

ID	Submission	Accession nb	Taxon	Sex	Country	Latitude	Longitude	Seq.Center	Mean RD	Raw reads	Mapped	Unmapped
c511	SRA106228	SRR1057584	*H. cydno galanthus*	Female	Costa Rica	10°16' N	84°11'W	BGI	14.57	56980695	54904440	2076255
c512	SRA106228	SRR1057585	*H. cydno galanthus*	Male	Costa Rica	9°40'N	83°2'W	BGI	14.67	57786331	55444921	2341410
c513	SRA106228	SRR1057586	*H. cydno galanthus*	Male	Costa Rica	10° 26'N	83° 59'W	BGI	14.83	58047481	56045188	2002293
c514	SRA106228	SRR1057587	*H. cydno galanthus*	Female	Costa Rica	9°43'N	83°3'W	BGI	14.73	57576892	55612600	1964292
c515	SRA106228	SRR1057588	*H. cydno galanthus*	Female	Costa Rica	10° 13'N	83° 41'W	BGI	14.70	57606552	55532956	2073596
c563	SRA106228	SRR1057589	*H. cydno galanthus*	Male	Costa Rica	10° 13'N	83° 47'W	BGI	13.83	54029777	52103112	1926665
c614	SRA106228	SRR1057590	*H. cydno galanthus*	Male	Costa Rica	9°43'N	83°3'W	BGI	14.81	57793383	55949422	1843961
c630	SRA106228	SRR1057591	*H. cydno galanthus*	Female	Costa Rica	10° 26'N	83° 59'W	BGI	13.83	54218576	52122166	2096410
c639	SRA106228	SRR1057592	*H. cydno galanthus*	Male	Costa Rica	10° 13'N	83° 41'W	BGI	14.84	57839669	55879865	1959804
c640	SRA106228	SRR1057593	*H. cydno galanthus*	Female	Costa Rica	9°43'N	83°3'W	BGI	14.82	57905245	55775583	2129662
m523	SRA106228	SRR1057594	*H. melpomene rosina*	Male	Costa Rica	9°43'N	83°3'W	BGI	14.94	58408944	55702153	2706791
m524	SRA106228	SRR1057595	*H. melpomene rosina*	Female	Costa Rica	9°51'0N	84°19'W	BGI	15.03	57327760	55904796	1422964
m525	SRA106228	SRR1057596	*H. melpomene rosina*	Male	Costa Rica	8°28'N	83°35'W	BGI	14.58	58361427	54333551	4027876
m589	SRA106228	SRR1057597	*H. melpomene rosina*	Male	Costa Rica	9°52'N	83°0'W	BGI	14.92	56991140	55771647	1219493
m675	SRA106228	SRR1057598	*H. melpomene rosina*	Male	Costa Rica	9°24'N	84°10'W	BGI	15.07	57387640	56029868	1357772
m676	SRA106228	SRR1057599	*H. melpomene rosina*	Male	Costa Rica	9°43'N	83°3'W	BGI	13.63	53924476	50738856	3185620
m682	SRA106228	SRR1057600	*H. melpomene rosina*	Male	Costa Rica	10°26'N	83°59'W	BGI	14.96	57455142	55639521	1815621
m683	SRA106228	SRR1057601	*H. melpomene rosina*	Male	Costa Rica	9°24'N	84°10'W	BGI	14.19	54047170	52772924	1274246
m687	SRA106228	SRR1057602	*H. melpomene rosina*	Female	Costa Rica	10°26'N	83°59'W	BGI	15.06	57500817	56116208	1384609
m689	SRA106228	SRR1057603	*H. melpomene rosina*	Female	Costa Rica	9°51'0N	84°19'W	BGI	14.71	57496048	54767875	2728173
247-1	ERA206886	ERR260277	*H. melpomene rosina*	Male	Panama	9°1206'N	79°6969'W	The GenePool	26.9	80357189	77576309	2780880
247-2	ERA206886	ERR260278	*H. melpomene rosina*	Male	Panama	9°1206'N	79°6969'W	The GenePool	26.7	80780396	77202364	3578032
247-3	ERA206886	ERR260279	*H. melpomene rosina*	Male	Panama	9°1206'N	79°6969'W	The GenePool	26.5	79483045	76412158	3070887
248-4	ERA206886	ERR260280	*H. melpomene rosina*	Male	Panama	9°1206'N	79°6969'W	The GenePool	36.7	109113325	1,07E+08	2578412
248-5	ERA206886	ERR260281	*melpomene melpome*	Male	French Guiana	4°9632'N	52°4200'W	The GenePool	24.1	89838758	71357196	18481562
248-6	ERA206886	ERR260282	*melpomene melpome*	Male	French Guiana	4°9632'N	52°4200'W	The GenePool	23.1	83950002	68504773	15445229
248-7	ERA206886	ERR260283	*melpomene melpome*	Male	French Guiana	4°9632'N	52°4200'W	The GenePool	35.0	107926916	1,04E+08	3643193
249-8	ERA206886	ERR260284	*melpomene melpome*	Male	French Guiana	4°9151'N	52°3755'W	The GenePool	35.8	110399158	1,06E+08	4072667
249-9	ERA206886	ERR260285	*melpomene melpome*	Female	Panama	8°6136'N	78°1398'W	The GenePool	62.0	190117929	1,85E+08	5488867
249-10	ERA206886	ERR260286	*melpomene melpome*	Male	Panama	8°2797'N	77°8098'W	The GenePool	15.6	47146484	44511911	2634573
252-19	ERA206886	ERR260295	*H. cydno chioneus*	Male	Panama	9°1714'N	79°7573'W	The GenePool	35.8	111609395	1,05E+08	6584033
252-20	ERA206886	ERR260296	*H. cydno chioneus*	Male	Panama	9°1714'N	79°7573'W	The GenePool	35.3	108838440	1,04E+08	5106924
252-21	ERA206886	ERR260297	*H. cydno chioneus*	Male	Panama	9°1714'N	79°7573'W	The GenePool	39.2	119944299	1,15E+08	4577836
252-22	ERA206886	ERR260298	*H. cydno chioneus*	Male	Panama	9°1714'N	79°7573'W	The GenePool	46.0	146341596	1,36E+08	10744005

Table S1. Illumina paired-end sample information

Illumina paired-end sequencing information for 20 *H. melpomene* and 14 *H. cydno* butterflies retrieved from public repositories for this study (SRA106228, Kronforst et al. 2013; ERP002440, Martin et al. 2013). *ID*, refers to the code ID given to each sample during this study; *Submission* and *Accession number* are as appear on the public repositories. *Taxon, Sex*, *Country, Latitude, Longitude* and *Seq. center* are as in the original publications. *Mean RD* (read-depth), (Total) *Raw reads*, *Mapped* (reads) and *Unmapped* (reads) values calculated after mapping to the *H. melpomene* genome (v2.0) (Davey et al. 2016) with Stampy (v1.0.23; Lunter & Goodson 2011) using default values for all parameters except the substitution rate, which was set to 0.01.

Percentage overlap	Number of duplications that overlap between both species
10	1482
20	1396
30	1386
40	1352
50	1380
60	1343
70	1280
80	1244
90	1107
91	1100
92	1059
93	1032
94	1010
95	983
96	945
97	895
98	619
99	407

Table S2. Number of duplications that overlap between both species decreases with greater overlap percentage thresholds

Total number of duplications calls that overlaps between the *H. melpomene* and the *H. cydno* depending on merging overlapping percentage criteria.

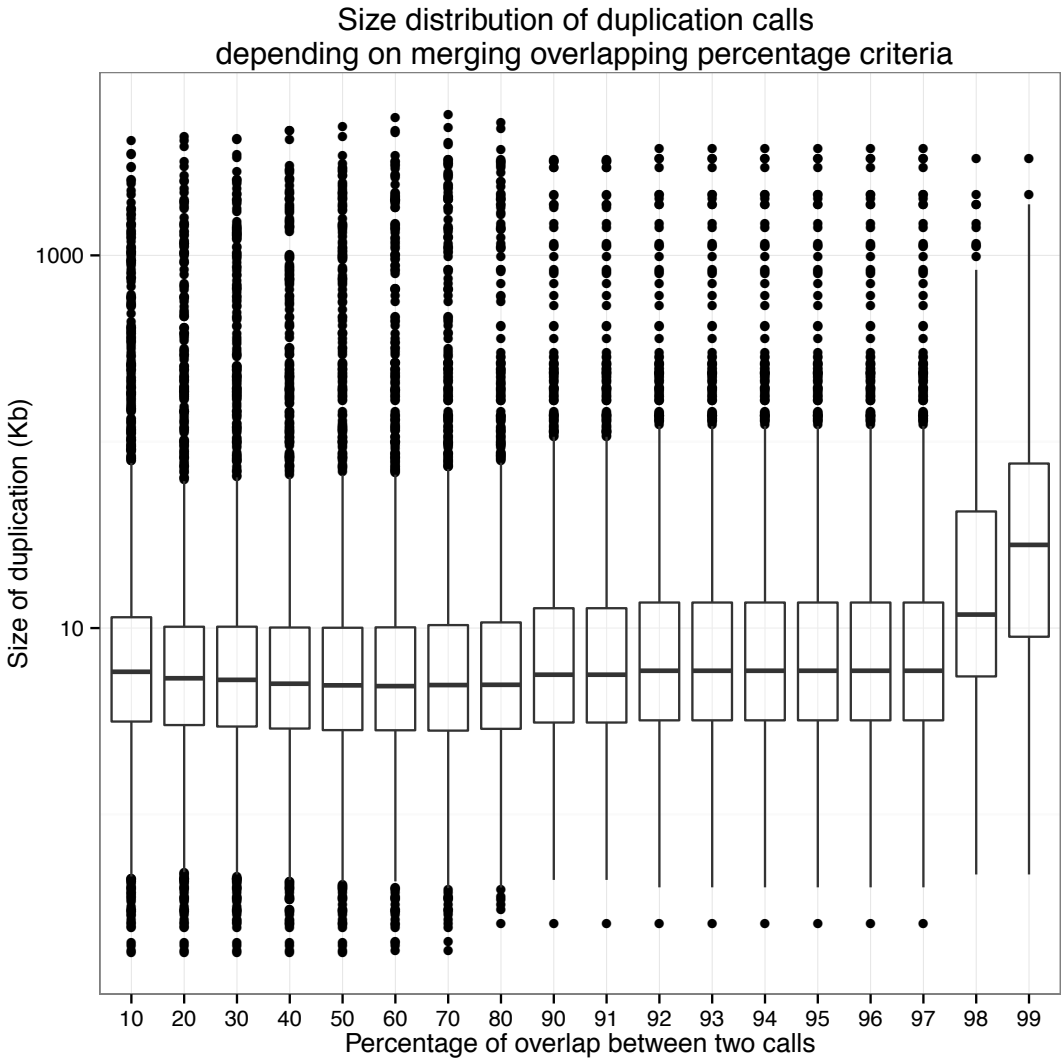
Chr	Scaffold	Start	End	Size	BayeScan log(PO)	BayPass mean XtX	Freq in *H. melpomene*	Freq in *H. cydno*	Hmel2 annotation
9	Hmel209007	4344840	4364959	20119	1.7222	7.95239143	0	0.93	HMEL009267
15	Hmel215006	1190144	1196212	6068	1.8414	8.78515118	0	1	intergenic (upstream of *cortex)*
18	Hmel218003	221730	429239	207509	1.894	8.75630075	0	1	OBP41 HMEL013558 HMEL013559 HMEL003174 HMEL003175 HMEL003862 HMEL003863
21	Hmel221012	779541	796444	16903	1.72	8.35788884	0	0.93	HMEL016617 HMEL016621 HMEL016620
2	Hmel202004	1537111	1543050	5939	Not Sig.	3.97886586	0.65	0.71	intergenic
21	Hmel221001	297653	480497	182844	Not Sig.	3.56709747	0.55	0.5	HMEL011045 HMEL011044 HMEL011042 HMEL011043 HMEL011041 HMEL011040 HMEL011037 HMEL011039 HMEL011038
2	Hmel202006	1791672	1816820	25148	1.5501	Not Sig.	0	0.79	HMEL015626
2	Hmel202006	1804029	1835608	31579	1.0762	Not Sig.	0.05	1	HMEL015625
2	Hmel202006	4165194	4171714	6520	0.99213	Not Sig.	0	0.93	HMEL015624
15	Hmel215047	2849864	2859210	9346	0.97681	Not Sig.	0.05	1	HMEL012374
19	Hmel219018	354635	361054	6419	1.2169	Not Sig.	0	1	HmGr58 HmGr59 HmGr60 HmGr61 HmGr64 HmGr65

Table S5. Duplications identified as outliers in the Heliconius Set

Position, size and summary statistics associated with BayeScan and BayPass. Frequency of the duplication in each species was calculated for all the 12 duplications. Statistics showing the significance of each call are shown for both BayeScan and BayPass. When the duplication event was not significant in one tool *Not sig.* was added to the column. Hmel2 annotations are shown for each duplication.

Supplementary Figures

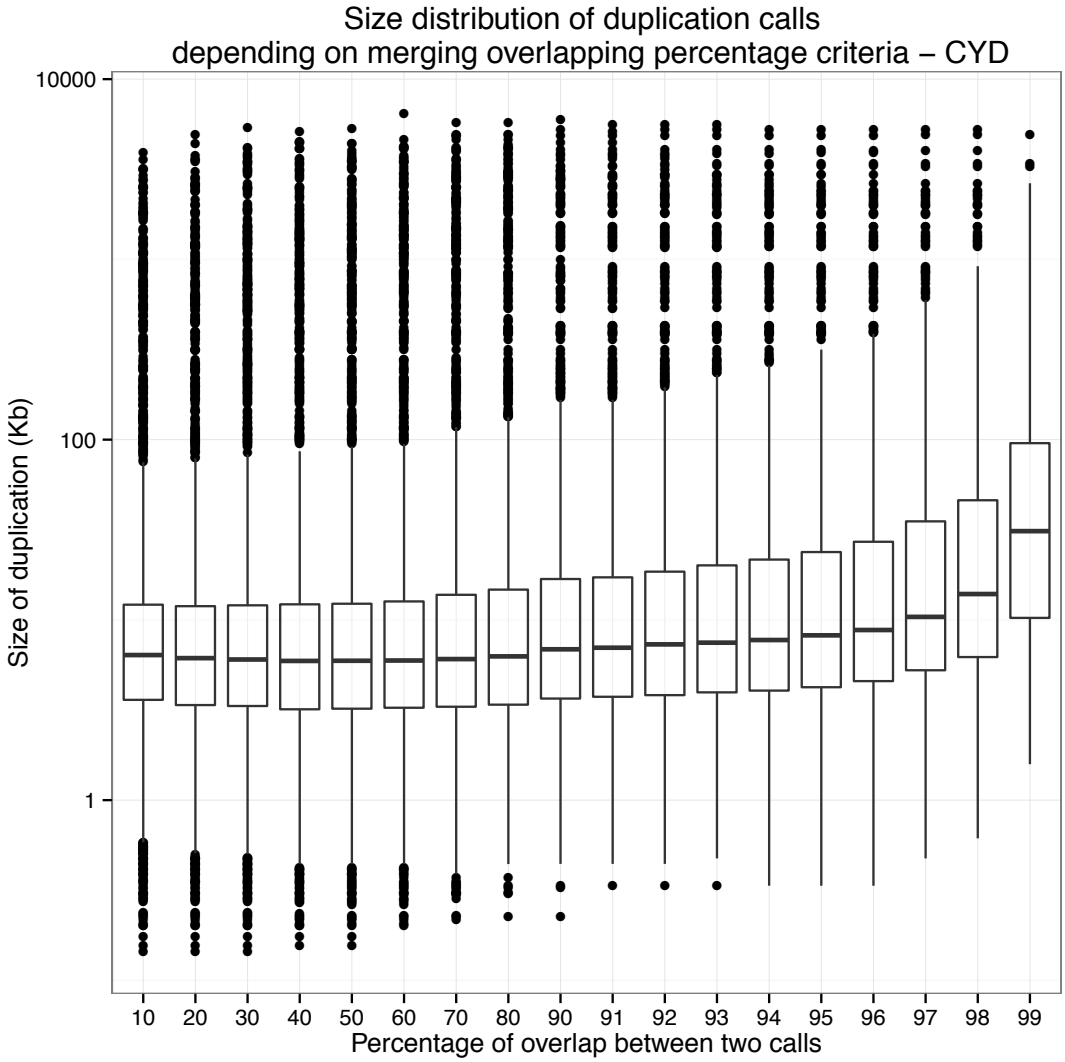
Supplementary Figure S1.



Supplementary Figure S1. Size Distribution of duplication calls depending on merging overlapping criteria for *H. melpomene*.

Depending on the chosen overlap percentage median sizes and distributions of duplications in the *H. melpomene* Discovery Set varies.

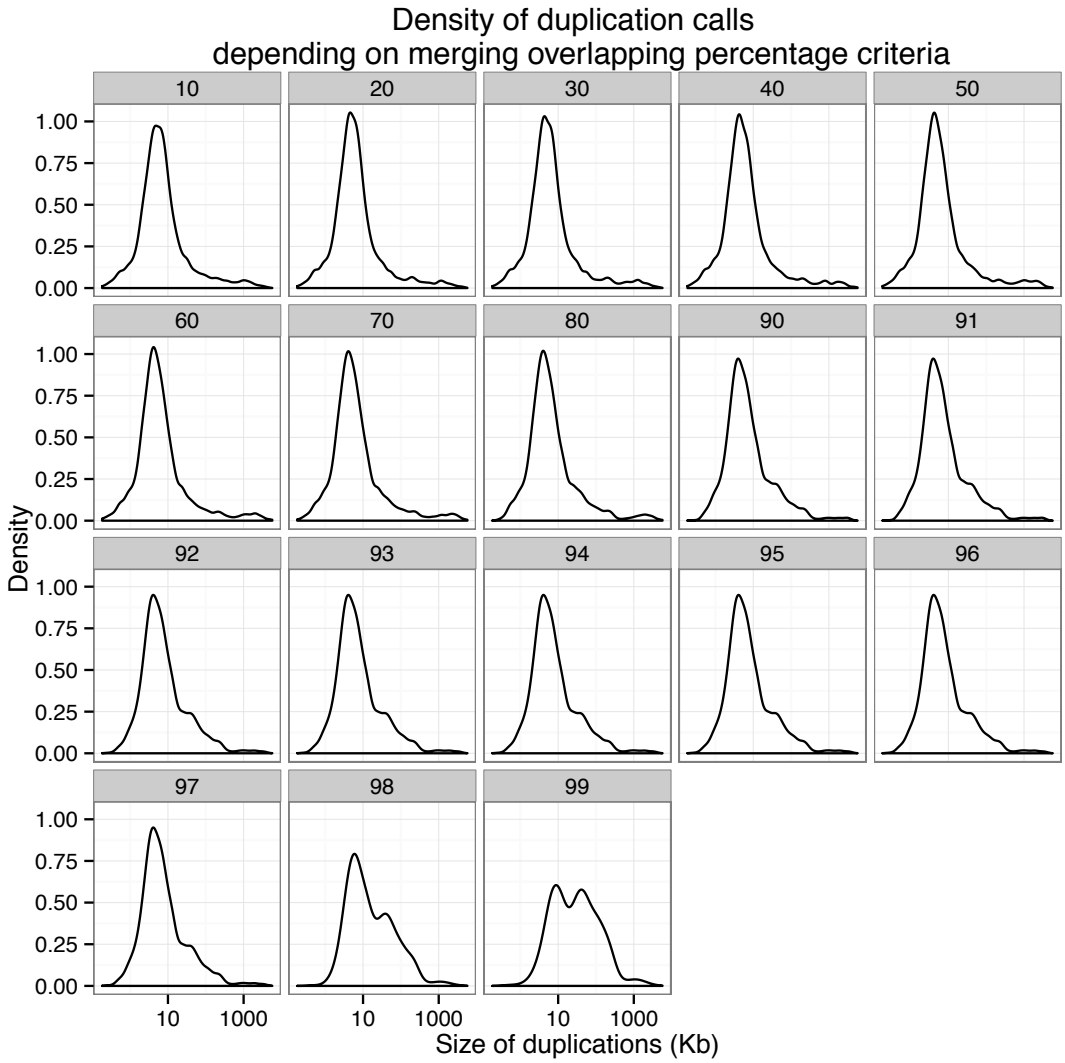
Supplementary Figure S2.



Supplementary Figure S2. Size Distribution of duplication calls depending on merging overlapping criteria for *H. cydno*.

Depending on the chosen overlap percentage median sizes and distributions of duplications in the *H. cydno* Discovery Set varies.

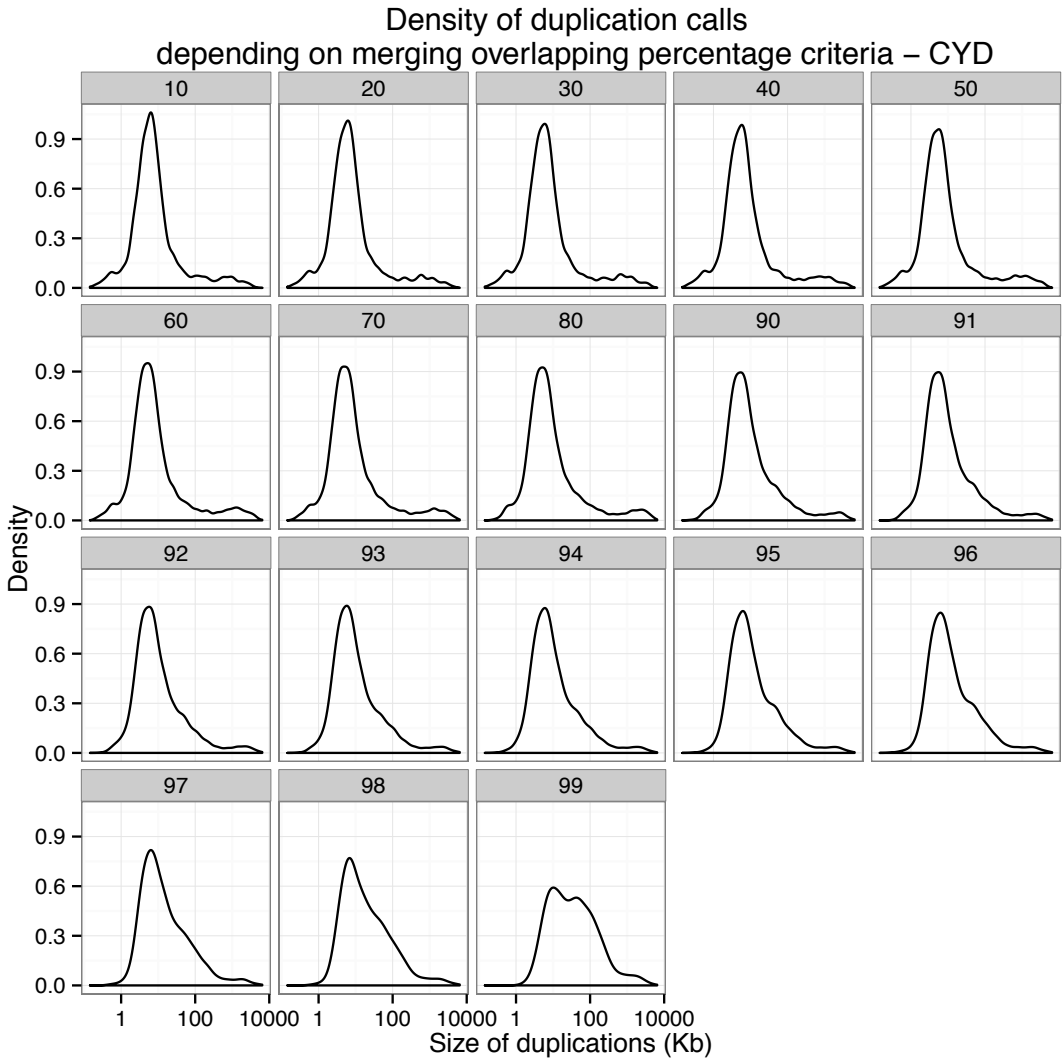
Supplementary Figure S3.



Supplementary Figure S3. Density distribution of duplication calls depending on merging overlapping criteria for *H. melpomene*.

Depending on the chosen overlap percentage median sizes and distributions of duplications in the *H. melpomene* Discovery Set varies.

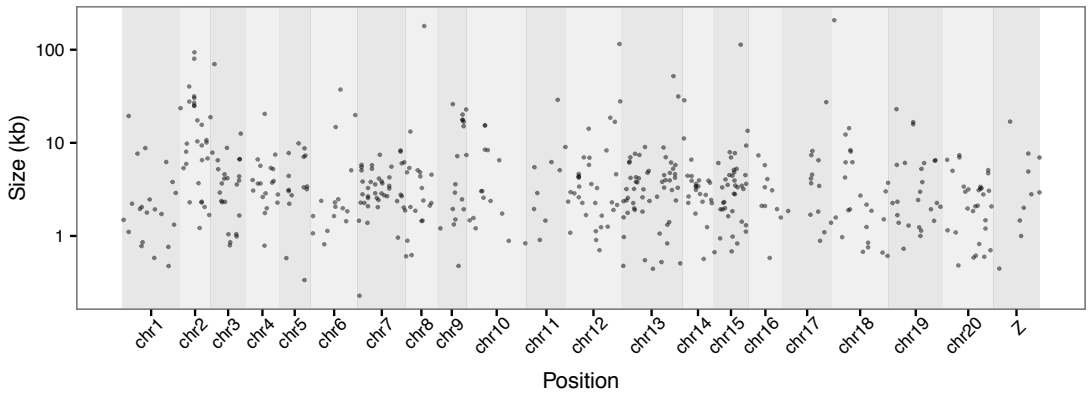
Supplementary Figure S4.



Supplementary Figure S4. Density distribution of duplication calls depending on merging overlapping criteria for *H. cydno*.

Depending on the chosen overlap percentage median sizes and distributions of duplications in the *H. cydno* Discovery Set varies.

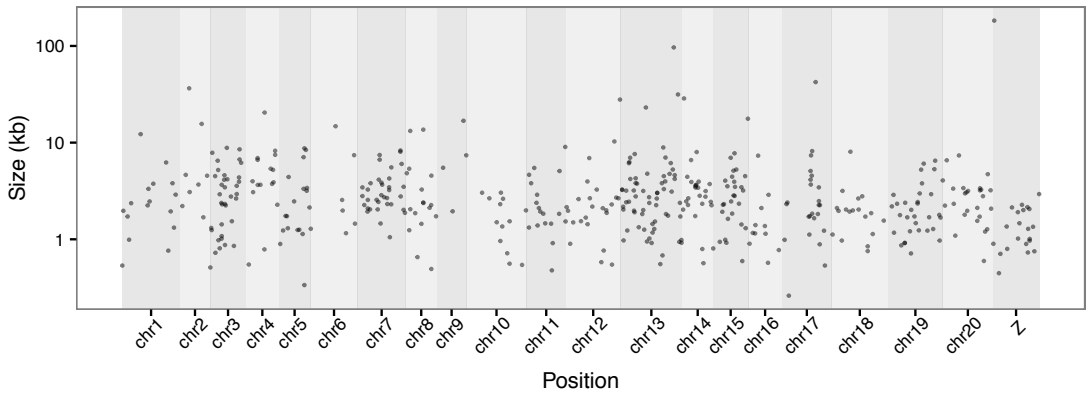
Supplementary Figure S5.



Supplementary Figure S5. Genotyping *H. cydno* set

Genotyped duplication set in *H. cydno* with 497 duplications

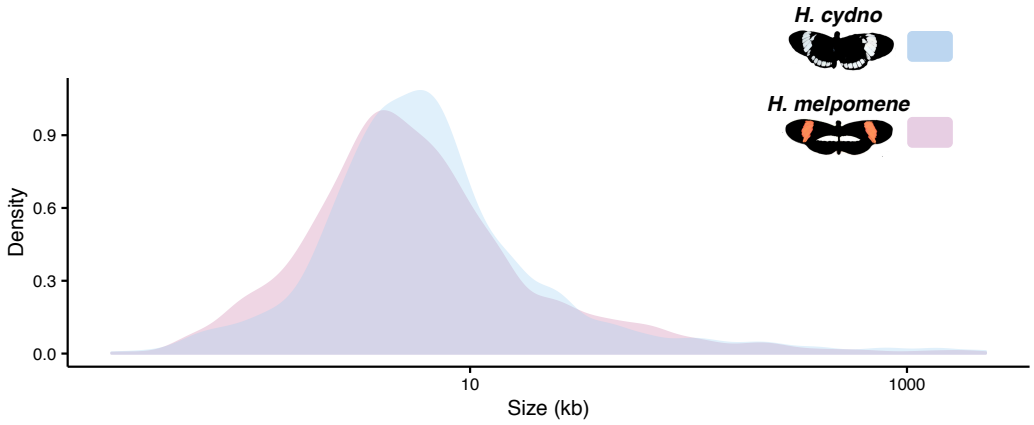
Supplementary Figure S6



Supplementary Figure S6. Genotyping *H. melpomene* set

Genotyped duplication set in *H. melpomene* with 463 duplications

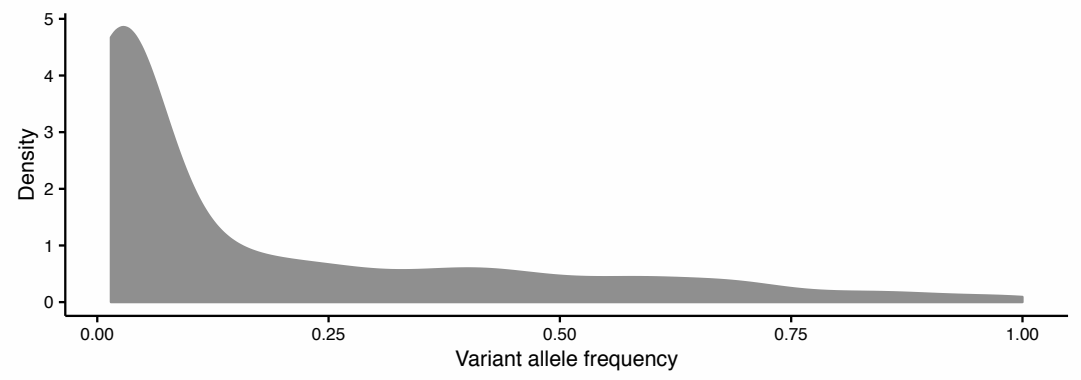
Supplementary Figure S7.



Supplementary Figure S7. Size distribution of the Genotyping *Heliconius cydno* and *H. melpomene* duplication sets

Size distribution of the calls for the *H. cydno* and *H. melpomene* Genotyping Sets. *H. cydno* is represented in blue and *H. melpomene* in pink. Size in kb.

Supplementary Figure S8.

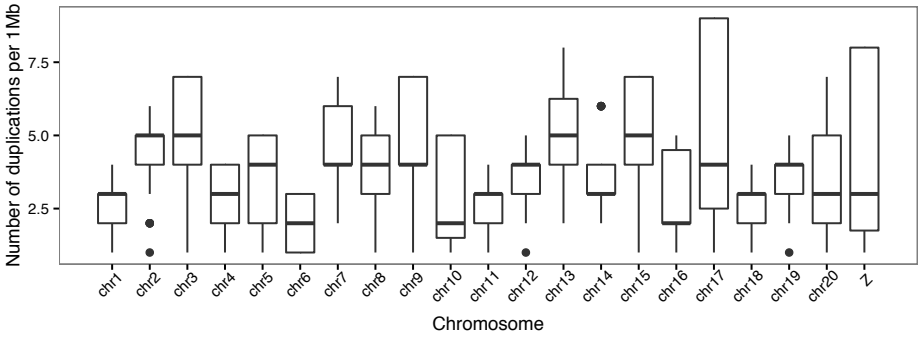


Supplementary Figure S8. Variant allele counts for the *H. melpomene* and *H. cydno* Genotyping sets and variant allele frequency in the Heliconius set

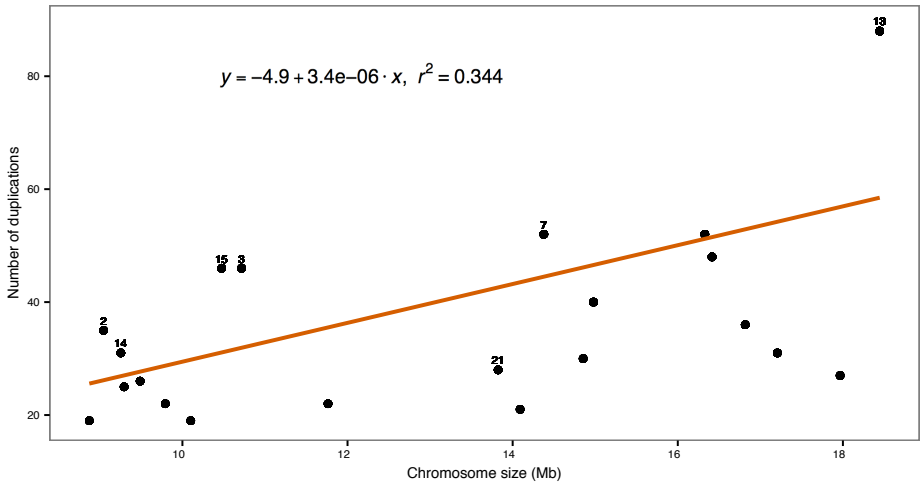
Variant allele frequency for the Heliconius Set for the 14 *H. cydno* and 20 *H. melpomene*. Duplication alleles also treated as co-dominant (presence/absence) markers.

Supplementary Figure S9.

A



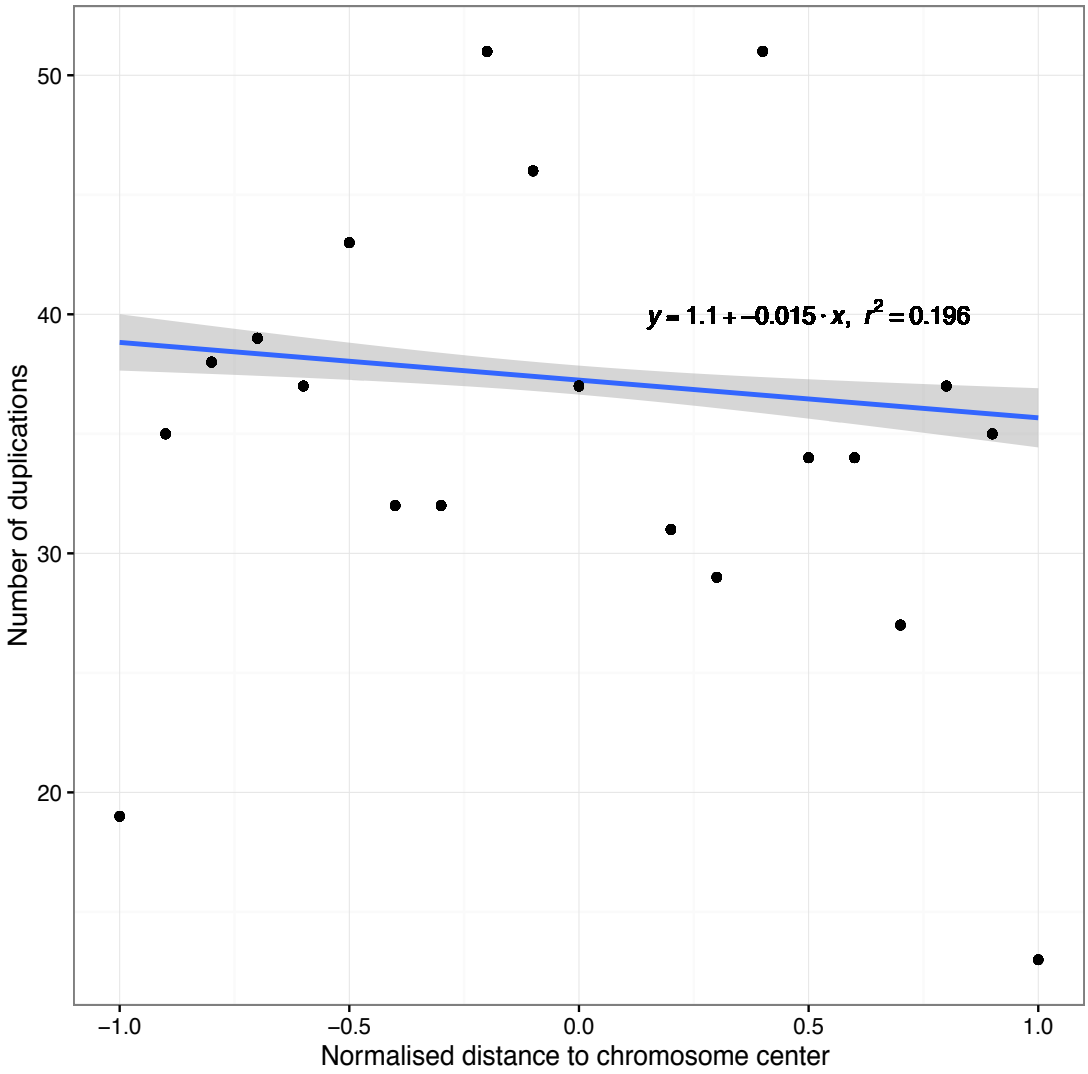
B



Supplementary Figure S9. Genome-wide distribution of duplications in the *Heliconius* set

A. Box-and-whisker plots displaying the number of inferred genotyped duplications per 1Mb-window for each chromosome. **B.** Overall the number of duplications genotyped in the *Heliconius* set correlates with chromosome size. Each point represents one chromosome. Chromosomes that have a greater number of absolute duplications than the fitted line are also identified by their number above the point. Chromosome 21 (Z, sex-chromosome) has also been identified.

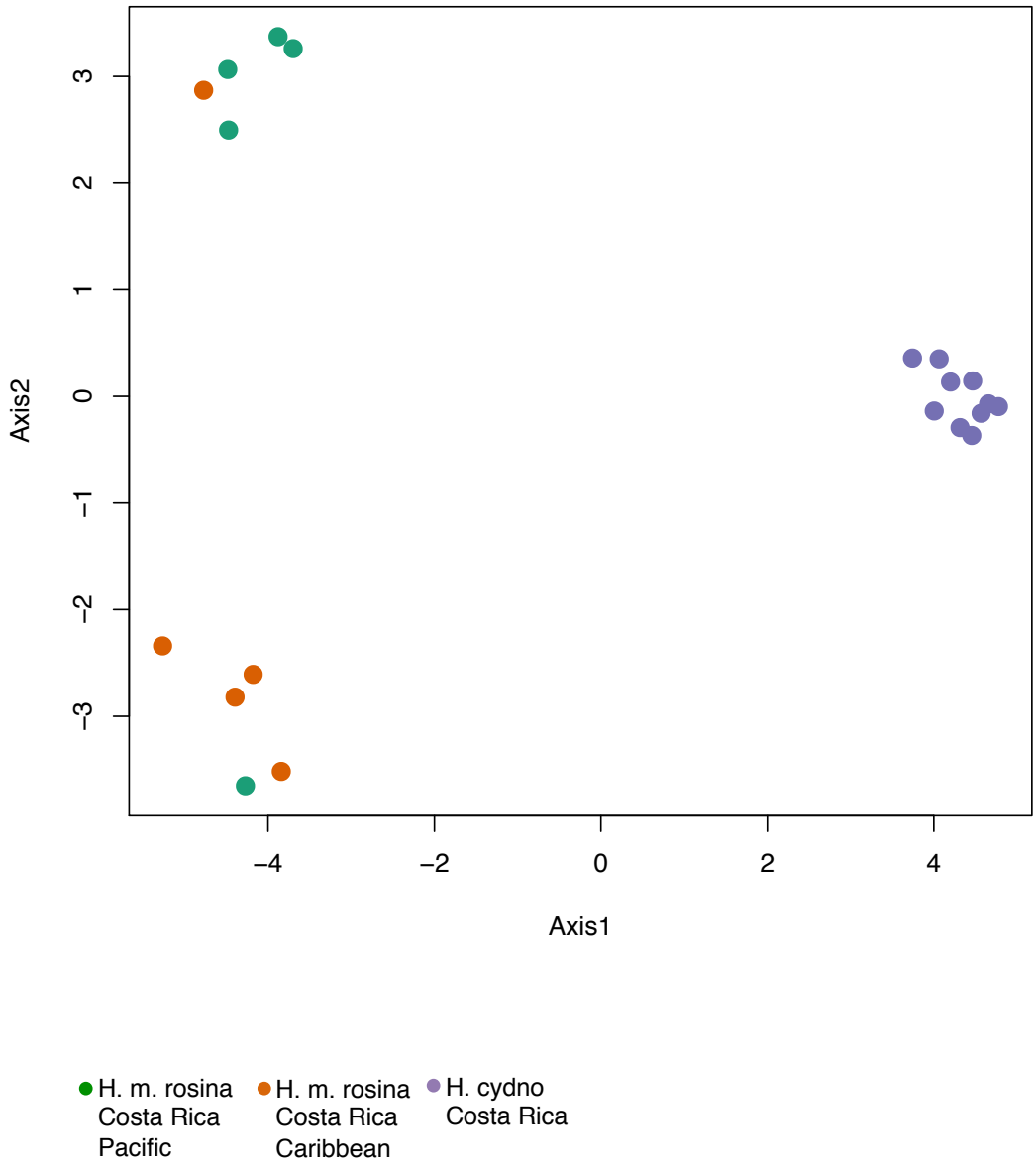
Supplementary Figure S10.



Supplementary Figure S10. Distribution of duplications in the Heliconius set along chromosome position

Number of duplications identified in the Heliconius set and their normalised distance from the chromosome centre. In the x axis -1 is the normalised chromosome position and 1 is the normalised chromosome end. 0 is the chromosome centre. Line fit to the correlation between normalised chromosome location and number of duplications.

Supplementary Figure S11.

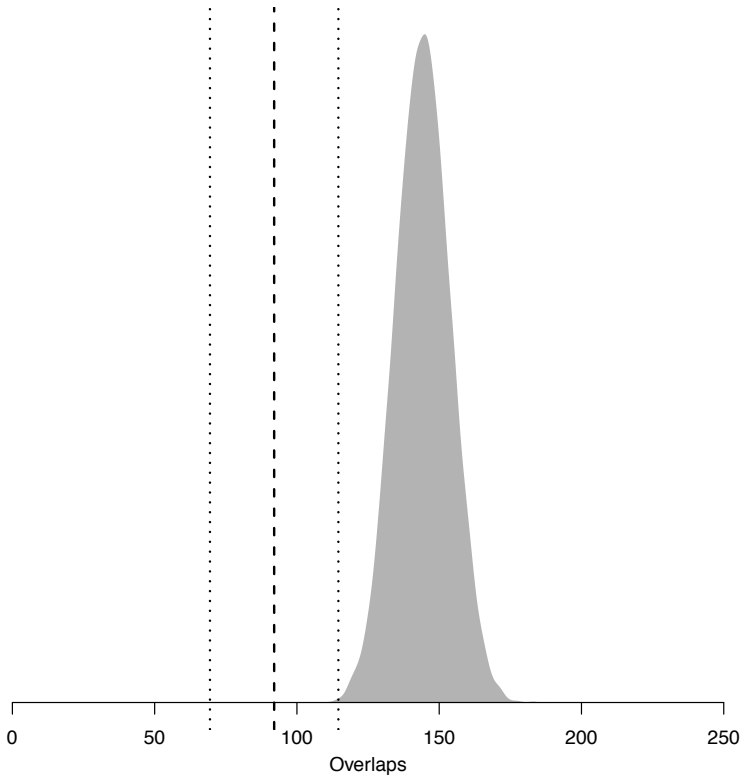


Supplementary Figure S11. Principal component analysis of the duplicated variants in the *Heliconius* set with Costa Rican samples

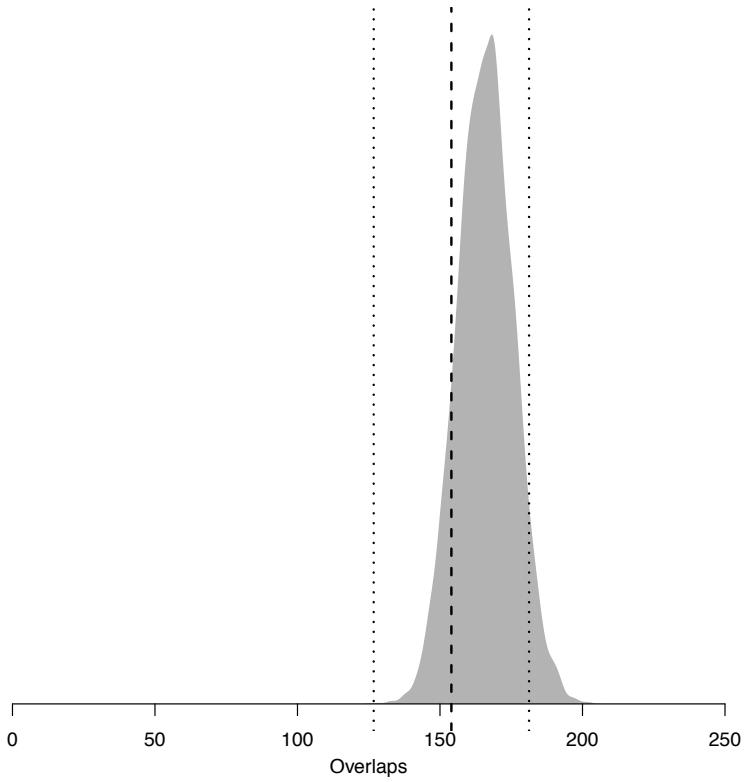
Samples cluster by species (PC1) and location (PC2) based on their duplication genotype. 23.47% of the total variance was explained by the first two principal components (PC1 18.856% and PC2 4.618%).

Supplementary Figure S12.

A.



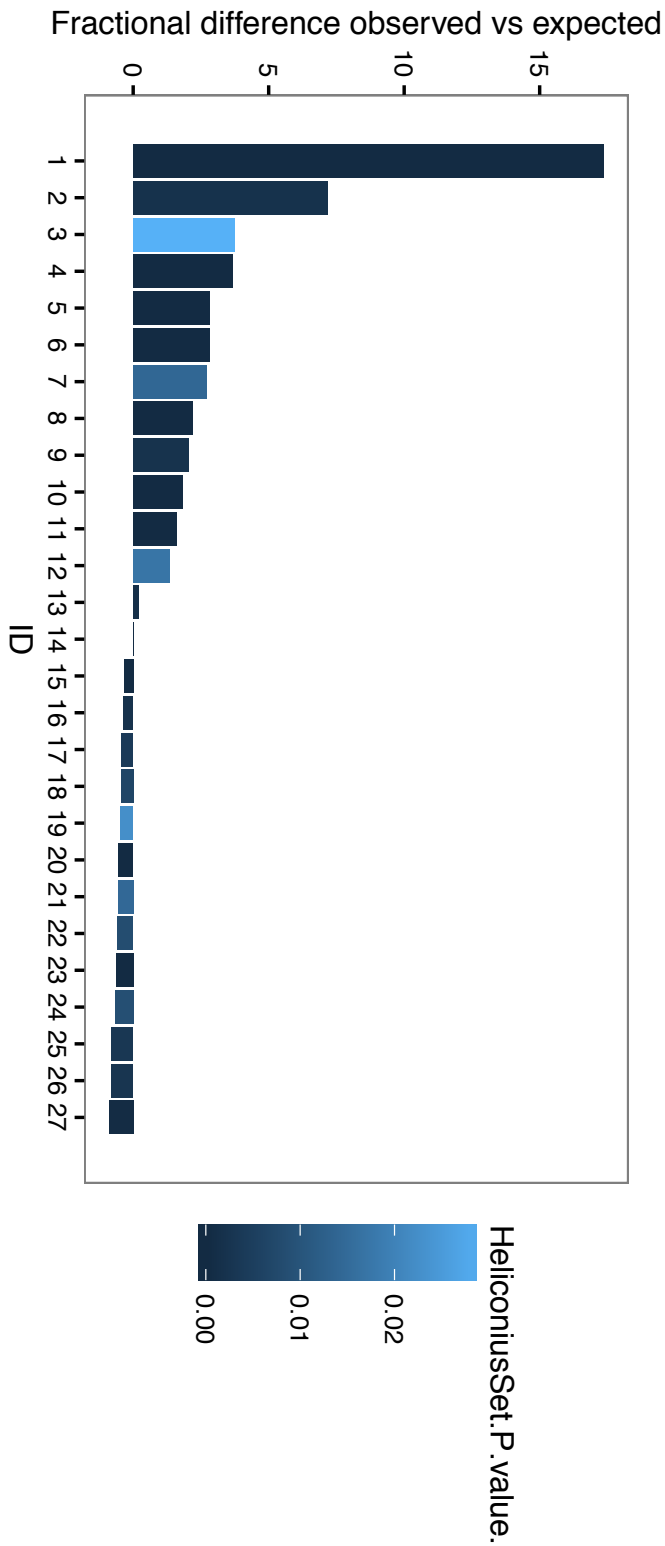
B.



Supplementary Figure S12. Observed and simulated proportions of duplications overlapping with coding regions

The grey shading represents the distribution of simulated overlaps on 10 000 random replicates for the proportion of sites overlapping coding regions for duplications in *H. melpomene* **A.** and *H. cydno* **B.** (Table 2, Gene %). Vertical dotted lines indicate the mean and standard deviation for the overlap from 10 000 random simulations. The vertical dashed line indicates the observed percentage overlap of duplications with genic sequence for *H. melpomene* **A.** and *H. cydno* **B.** (Table 2, Gene %).

Supplementary Figure S13.



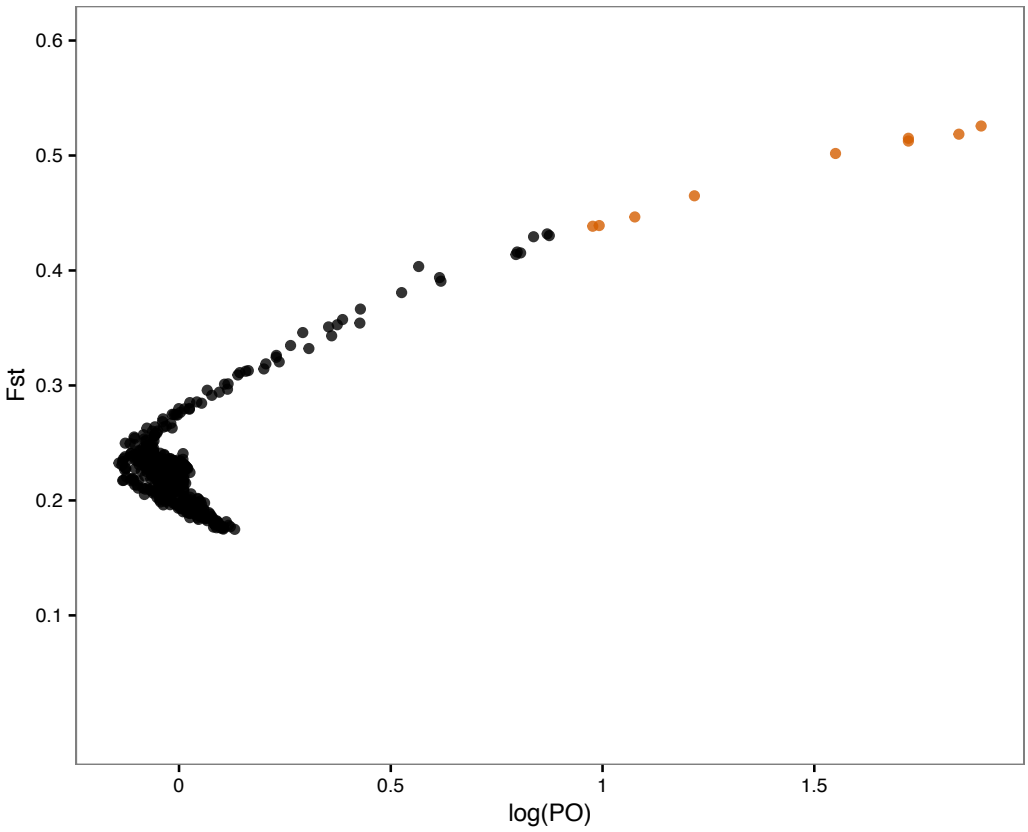
ID	PANTHER GO-Slim Biological Process
1	pentose-phosphate shunt (GO:0006098)
2	peroxisomal transport (GO:0043574)
3	vitamin biosynthetic process (GO:0009110)
4	cellular amino acid biosynthetic process (GO:0008652)
5	respiratory electron transport chain (GO:0022904)
6	DNA replication (GO:0006260)
7	vitamin transport (GO:0051180)
8	generation of precursor metabolites and energy (GO:0006091)
9	protein glycosylation (GO:0006486)
10	DNA metabolic process (GO:0006259)
11	proteolysis (GO:0006508)
12	steroid metabolic process (GO:0008202)
13	primary metabolic process (GO:0044238)
14	Unclassified (UNCLASSIFIED)
15	cellular process (GO:0009987)
16	biological regulation (GO:0065007)
17	RNA metabolic process (GO:0016070)
18	cell communication (GO:0007154)
19	transcription, DNA-dependent (GO:0006351)
20	regulation of biological process (GO:0050789)
21	developmental process (GO:0032502)
22	phosphate-containing compound metabolic process (GO:0006796)
23	response to stimulus (GO:0050896)
24	translation (GO:0006412)
25	response to stress (GO:0006950)
26	immune system process (GO:0002376)
27	protein phosphorylation (GO:0006468)

Supplementary Figure S13. Enrichment of certain biological process classes in the Heliconius Set against the *D. melanogaster* PANTHER reference

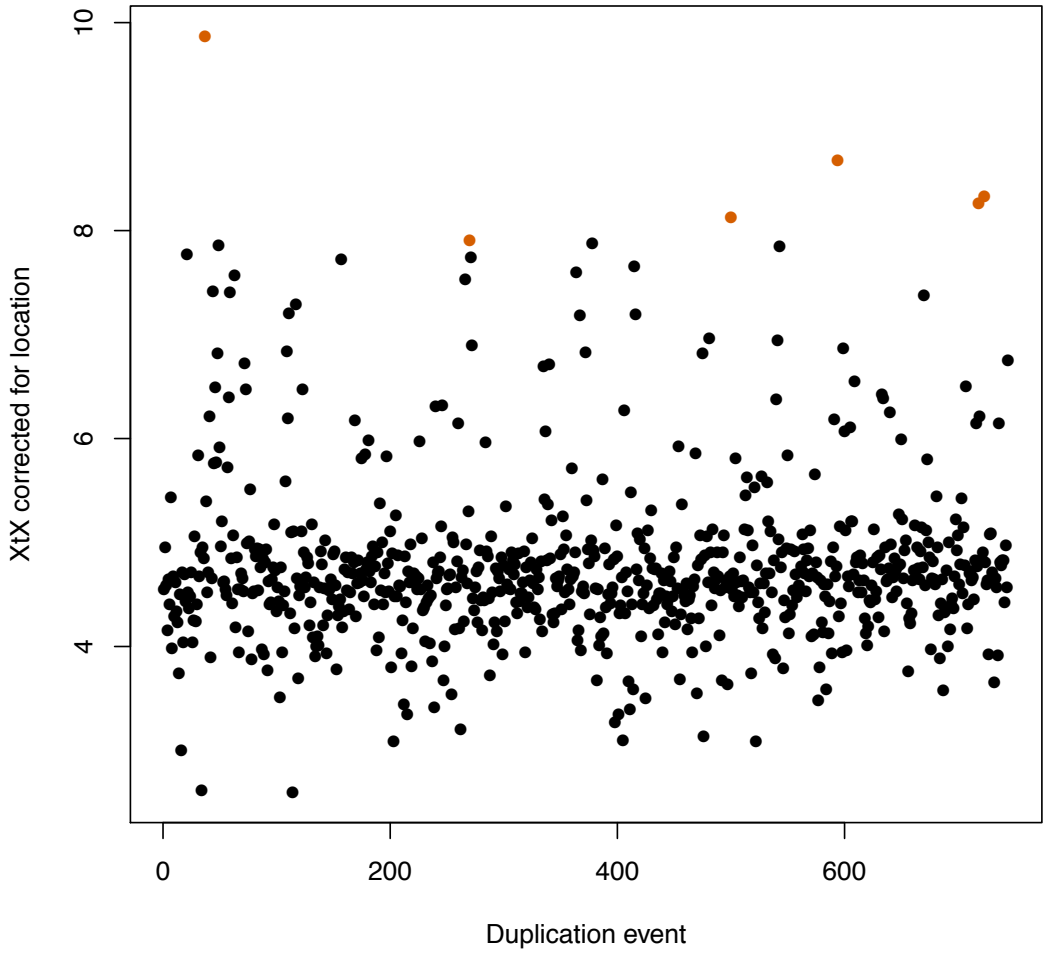
Fractional differences on the number of genes observed against the expected for each biological process category. Difference was calculate as the (number of genes observed for the category – number of gene expected for the category) / number gene expected for the category. Bars coloured by P value associated with each call, where $P<0.05$. Biological categories were given an ID from 1 to 27 from the one with the highest fold change to the one with the least. Negative values indicate a significant depletion genes associated with the biological process in question. Positive values an enrichment. PANTHER GO-slim Biological Process categories and the GO term associated with them shown below the fractional difference analysis.

Supplementary Figure S14.

A.



B.


XtX corrected for location
10
8
6
4
0
200
400
600
Duplication event

Supplementary Figure S14. Scan for selection on genomic regions putatively duplicated

A. BayeScan outlier analysis on the Heliconius duplication set using 14 *H. cydno* and 20 *H. melpomene* samples genotyped as dominant markers. 9 out of the 744 identified as putative duplications are expected to be under divergent selection between *H. cydno* and *H. melpomene* (FDR<0.05, shown in orange). log(PO), posterior odds score. **B.** BayPass outlier analysis on the Heliconius Set as dominant markers. Analysis was performed using the following covariates Costa Rica: 9.7489, 83.7534; Panama: 8.5380, 80.7821; French Guiana: 3.9339, 53.1258; *H.* cydno: 1 and H*. melpomene*: 2. Horizontal line represents the 98% threshold of the simulated data (XtX>7.9). Points above the threshold correspond to those duplications regions identified by the outlier analysis after correcting for location. x axis plots each duplication in the Heliconius set (744 duplications in total). y axis represents the mean XtX for the region.

CHAPTER 2

Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*

Abstract

Heteromorphic sex chromosomes follow different evolutionary trajectories as compared to autosomes, reflecting distinct evolutionary pressures. The strength of positive selection, genetic drift and purifying selection differs between sex chromosomes and autosomes resulting in different rates of molecular evolution. In particular, recessive mutations are exposed to selection more readily on the sex chromosomes in the heterogametic sex. Using publicly available male and female whole-abdomen transcriptome data for *Heliconius melpomene*; newly generated female transcriptome data from ovary and gut tissue; and WGS *H. melpomene* and *H. erato* data; I measure the strength of positive and negative selection and rates of adaptive evolution between the Z- and autosomal-linked genes. I show that positive selection is higher in Z-linked female-biased genes. In *Heliconius* hemizygosity might affect the rate of adaptive substitutions; but there is no significant difference in rate of adaptive evolution, positive or purifying selection between ovary-biased and gut-biased genes. Together these results do not support a fast-Z effect or a reduced efficacy of purifying selection in Z-linked genes, despite a low effective population size of the Z chromosome. The lack of a fast-Z effect in *Heliconius* adds to a growing body of literature from other ZW systems without a global dosage compensation mechanism that also lack a fast-Z effect.

Introduction

Sexually dimorphic expression is often caused by natural and/or sexual selection favouring phenotypes that influence the fitness of one of the sexes. In species with genetic sex determination males and females are almost genetically identical despite the observed phenotypic differences between the sexes. Male and female genomes differ by a few genes usually located on non-recombining sex-specific regions of the genome. Therefore, the majority of sexually dimorphic traits result from the differential expression of genes present in both male and female genomes (Ellegren and Parsch, 2007).

Genes with sex-biased expression are common in all taxa that have been studied: from mammals to diptera, reptiles, birds and lepidoptera (Rinn and Snyder, 2005; Mank, Nam, *et al.*, 2010). For example, when whole *Drosophila melanogaster* adult female and male gene expression landscapes were compared, 57% of genes were categorised as sex-biased (Assis *et al.*, 2012). For *Heliconius melpomene*, 13% of genes were categorised as male-biased and 16% as female-biased when expression patterns were compared between male and female whole-abdomen and heads (Walters *et al.*, 2015).

The vast majority of genes that exhibit sexually dimorphic expression are expressed in reproductive tissues and, with different patterns of expression in males and females, sex-biased genes tend to also have distinctive rates of molecular evolution (Parisi *et al.*, 2003; 2004; Avila *et al.*, 2015). Comparison of non-synonymous substitutions to synonymous substitutions, dN/dS, have shown that sex-biased genes tend to diverge faster than the genome average, and that in XY systems, male-biased genes are the most divergent between species (Kirkpatrick and Hall, 2004; Zhang *et al.*, 2004; Assis *et al.*, 2012; Nam *et al.*, 2015). Therefore, the identification of sex-biased genes and subsequent analysis of patterns of molecular evolution, ultimately contributes to a better understanding of the evolutionary forces shaping sex chromosome and autosome evolution.

In addition to patterns of gene expression, sex also influences molecular evolution trough the patterns of inheritance at sex chromosomes (Rice, 1984). Population genetic theory predicts that sex chromosomes may have a disproportionate role in the evolution of divergence and, subsequently, in the process of speciation. Ancient sex chromosomes are effectively haploid in one sex – males in XY systems and females in ZW systems. This is may result in an increased evolutionary rate of sex chromosomes relative to autosomes, a phenomenon known as the fast-X effect (Charlesworth *et al.,* 1987). Driven in part by their pattern of inheritance, X-linked genes can diverge faster between species than autosomal-linked genes if certain parameters of: 1) allelic dominance; 2) selection in males versus females; 3) mutation; 4) recombination and 5) effective population size (Ne), are met (Orr and Betancourt, 2001; Kirkpatrick and Hall, 2004; Vicoso and Charlesworth, 2006; 2009; Orr, 2010; Connallon *et al*., 2012). The analysis of divergence rates between sex-linked and autosomal-linked genes, however, has produced mixed evidence in support of the fast-X effect (Meisel and Connallon, 2013).

Faster-X evolution studies measure two different metrics: 1) dN/dS; and 2) the amount of adaptive evolution (α) using the McDonald-Kreitman test (McDonald and Kreitman, 1991). Studies measuring dN/dS are testing for "faster-X divergence" and, although useful for comparing X-linked versus autosomal-linked divergence, capture the effects of both adaptive, and neutral or slightly deleterious mutations. Estimates of α, combine measures of within species polymorphism and between-species divergence, and test for "faster-X adaptation". In some taxa there is strong evidence for faster-X divergence but not faster-X adaptation and vice versa (Meisel and Connallon, 2013). For example, the first calculations for faster-X divergence were carried out in *Drosophila* where support for elevated dN/dS in X-linked genes has been mixed. Studies that used autosome-to-X translocations to control for gene content effect did not reach consensus on the existence of faster-X divergence (Thornton *et al*., 2006; Zhou and Bachtrog, 2012), but X-linked duplicate genes have elevated dN/dS compared to autosomal duplicates

(Thornton and Long, 2002). Signals of faster-X divergence in *Drosophila* have been shown to exist at non-coding sites which could reflect a higher mutation rate on the X chromosome compared to autosomes or the fixation of recessive advantageous substitutions that affect genes in *cis* (Hu *et al.*, 2013). However, faster-Z divergence tests in other taxa has had stronger support. For example in humans, chimpanzees and rodents dN/dS is higher for X-linked genes (Nielsen *et al.*, 2005; Mank, Vicoso, *et al.*, 2010). In birds, a ZW sex determination system, the existence of faster-Z divergence has been reported but Z-linked male-biased genes were not less accelerated than unbiased genes or female-biased genes (Wright *et al.*, 2015). This is not expected if the fast-Z effect is driven by recessive beneficial mutations and so does not reflect positive selection (Mank, Nam, *et al.*, 2010).

On the other hand, whole-genome analyses have resulted in stronger evidence for higher frequencies of adaptive substitutions among *Drosophila* X-linked genes (faster-X adaptation) (Mackay *et al.*, 2012). However, support for faster-X adaptation in vertebrates is less clear. McDonald-Kreitman tests support a faster-X adaptation for wild mouse populations but, for the European rabbit (*Oryctolagus cuniculus*), a clear faster-X adaptation signal is only present in populations with large effective population sizes (Baines and Harr, 2007; Carneiro *et al.*, 2012). In a recent study on satyrine butterflies, the authors also did not find significant differences in adaptive evolutionary rates between the Z and the autosomes (no faster-Z adaptation). However, the comparison of male-biased, female-biased and unbiased Z-linked genes revealed increased purifying selection against recessive deleterious mutations in female-biased Z-linked genes (Rousselle *et al.*, 2016).

Here I address these questions in *Heliconius melpomene*, a neotropical species of Lepidoptera with a ZW sex determination system. Previous analysis of *Heliconius* transcriptome data focused on the evolution of dosage compensation and the impact of sex-specific dosage on the levels of gene expression (Walters *et al.*, 2015). First, using the same transcriptome data, I briefly revisit this topic using a more complete *H. melpomene* reference

annotation. Second, I compare the strength of positive and purifying selection between the Z and autosomes accounting for sex-biased gene expression; and the rate of adaptive evolution between sex-biased Z- and autosomal-linked genes. Finally, I analyse newly generated female transcriptome data from ovary and gut tissue. Using this data I compare the strength of positive and purifying selection between germline tissue and somatic tissue to investigate whether genes expressed in the reproductive tissue of the heterogametic sex have higher rates of adaptive evolution than those expressed in somatic tissue.

Material and Methods

Samples

Gene expression data was calculated from: 1) 100bp paired-end mRNA-seq data from 5 *H. m. rosina* whole-male abdomens, and 5 *H. m. rosina* whole-female abdomens, downloaded from GenBank (BioProject PRJNA283415) (Walters *et al.*, 2015) with NCBI SRA toolkit (v2.5.7; National Center for Biotechnology Information, Bethesda, MD, USA); and 2) newly sequenced 150bp paired-end directional mRNA-seq data from ovary tissue of 7 young and 6 old *H. m. rosina* females, and from gut tissue of 6 young and 6 old *H. m. rosina* females (25 samples from 13 different individuals, Supplementary Table S1).

For these 25 samples *H. m. rosina* females were reared in insectaries in Gamboa, Panama. *P. triloba* potted plants were monitored daily and 5th instar caterpillars were removed and taken to the laboratory in large individual containers where they were allowed to pupate and emerge at a constant temperature (24-25ºC). The pupating containers in the laboratory were monitored several times a day for eclosion. When a female eclosed it was either: 1) taken back to the insectaries to be mated to a *H. m. rosina* male (Treatment: old, Supplementary Table S1); 2) or it was dissected 1h after

eclosion under controlled laboratory conditions (Treatment: young, Supplementary Table S1). Mated females were kept in individual 1m x 1m x 2m cages for 20 days until dissection.

Guts and ovaries were dissected in RNAlater at 24-25ºC RNAlater (ThermoFisher, Waltham, MA); and tissue was stored in RNAlater at 4ºC for 24h and -20ºC thereafter (ThermoFisher, Waltham, MA). Total RNA was extracted with a combined guanidium thiocyanate-phenol-chloroform and silica matrix protocol using TRIzol (Invitrogen, Carlsbad, CA), RNeasy columns (Qiagen, Valencia, CA) and DNaseI (Ambion, Naugatuck, CT) (Appendix B, Protocol for dissections of the reproductive tract for total RNA extraction). mRNA isolated from total RNA via poly-A pull-down, directional cDNA libraries and 150bp PE sequencing by Novogene Bioinfomatics Technologies (Hong Kong, China) (Supplementary Table S1).

Update Hmel2 released annotation

Hmel2 gene predictions are Hmel1 liftovers by the authors (The Heliconius Genome Consortium, 2012; Davey et al., 2016). The Hmel2 annotation file has 13 178 predicted transcripts spanning 16 897 139 bp. The Hmel2 annotation file is likely an under-representation of the existing features. For example, there are 20 118 high quality predicted transcripts in H. erato spanning 33 669 374 bp (van Belleghem et al., 2017). To improve the completeness of the annotation for *H. melpomene* I downloaded RNAseq reads from NCBI repositories ArrayExpress ID: E-TAB-1500 (Briscoe et al., 2013), and BioProject PRJNA283415 (Walters et al., 2015), published since Hmel1 release. I also used unpublished data from 10 wing RNAseq libraries (wing data generated by Joe Hanly). In collaboration with Sujai Kumar, the BRAKER1 pipeline was used to perform unsupervised RNA-seq based genome annotation. GeneMark-ET was used to perform iterative training, generating initial gene structures and AUGUSTUS was used for training and subsequent integration of RNAseq read information into the final gene

predictions (Stanke et al., 2008; Lomsadze et al., 2014; Hoff et al., 2016). This resulted in 26 017 predicted transcripts spanning 32 222 367 bp. 6 532 of these transcripts were considered repeat proteins based on 90% single hit match to repeat databases and were removed. We transferred the 428 manually annotated genes (441 transcripts/protein) from the original Hmel2 annotation and removed any BRAKER1 predictions that overlapped. We also transferred 189 genes (189 transcripts/proteins) that have been manually annotated and published since Hmel2 release. Specifically, we transferred 73 gustatory receptors; 31 immune response and 85 Glutathione-S-transferases and Glucuronosyltransferases (Briscoe et al., 2013; van Schooten et al., 2016; Yu et al., 2016) and removed any BRAKER1 predictions that were overlapping. This resulted in an annotation file with 20 102 genes (21 661 transcripts/proteins) (*H. melpomene* annotation files available from https://www.dropbox.com/sh/5krc7kn3u0oviwj/AADHTlQsoxQCnqZnivatNdRba?dl=0). BRAKER1 predictions that had 1-to-1 overlaps with Hmel1 names were replaced by their original Hmel2 name. For many-to-1 mapping between the BRAKER1 predictions and Hmel2, Hmel2 names were reused and a suffix of g1/g2/g3/etc was added. The rest are renamed from HMEL030000 onwards.

Read mapping, counting and estimation of variance-mean dependence

HISAT2 (Kim *et al.*, 2015) was used to align fastq reads to gene sequences from *H. melpomene* annotation file (*H. melpomene* annotation files available from https://www.dropbox.com/sh/5krc7kn3u0oviwj/AADHTlQsoxQCnqZnivatNdRba?dl=0) using default mapping parameters. Summary mapping statistics were calculated using samtools flagstat (v1.2) (Li *et al.*, 2009). htseq-count was used to count how many aligned sequencing reads mapped to each genic feature (HTSeq v0.6.1; python v2.7.10; option: -m union) (Anders *et al.*, 2015).

Estimation of variance-mean dependence from the count data was performed with the DESeq2 (v1.14.1) of Bioconductor (v3.4) in the R software environment (v3.2.5) using the constructor function DESeqDataSetFromHTSeqCount(design=~batch+sex) for sex-biased genes; and DESeqDataSetFromHTSeqCount(design=~batch+tissue) for ovary and gut-biased genes. All the result tables were built using the DESeq2 results() function (options: betaPrior=false, test=Wald) (Love *et al.*, 2014). I filtered the results as in Walters *et al.* (2015) with log2 fold significance threshold > |1.5| and FDR < 0.05 (options: lcfThreshold=1.5, altHypothesis="greaterAbs", alpha=0.05) (Walters *et al.*, 2015).

Identification of sex-biased genes and ovary- and gut-biased genes

Sex-biased genes are genes with sexually dimorphic expression. These genes include those that are expressed 1) just in one sex (sex-specific expression) and, 2) in both sexes but at a higher level in one sex (sex-enriched expression). Sex-biased genes can be further separated into male-biased and female-biased depending on which sex shows higher expression. To identify sex-biased genes in *H. melpomene* I used the 5 whole male and 5 whole female abdomens (Walters *et al.*, 2015). Ovary- and gut-biased genes include those that are expressed in 1) just the ovary tissue, or just the gut tissue; and 2) in both ovary and gut tissues but at a higher level in one of the two. To identify ovary and gut biased genes I used the 25 mRNA-seq samples generated for this project.

Extraction of orthologous genes and coding sequence alignment

OrthoFinder was used to identify orthologous groups of genes in the *H. melpomene* and the *H. erato* transcriptomes (options: -t 48 -a 6). 1-1

orthologous gene sequences between the two species were selected for use in subsequent analysis (Supplementary Table S2). Using Gff-Ex, a genome feature extraction package (Rastogi and Gupta, 2014), I extracted: 1) coding sequences from 10 whole-genome short-read re-sequenced wild *H. m. rosina* from Panama (Supplementary Table S3) mapped to Hmel2 (Davey *et al.*, 2016) with bwa (Li and Durbin, 2009); 2) coding sequences from the reference *H. erato* genome (van Belleghem *et al.*, 2017).

For the 10 whole-genome re-sequence *H. m. rosina* samples, variants were called using HaplotypeCaller (GATK v3.4-0-g7e26428) (options DP=8) (DePristo *et al.*, 2011). The coding fasta sequences from 1) and 2) corresponding to 1-1 orthologous genes in *H. melpomene* and *H. erato* were aligned using MACSE accounting for frameshifts and stop codons (Ranwez *et al.*, 2011).

π_S/π_n and dNdS ratios influence on expression level

To test whether gene expression level and chromosome type have a significant effect on π_S/π_n and dNdS ratios I used a multiple regression analysis. I establish the linear models:

$$\log(\pi_{nij}) \sim \log(\pi_{sij}) + \text{chromosome\_type}_j + \log(\text{FPKM}_i)$$

$$\log(d_{Nij}) \sim \log(d_{sij}) + \text{chromosome\_type}_j + \log(\text{FPKM}_i)$$

using R (v3.2.5). FPKM_i is the mean FPKM of gene *i* across the 10 individuals. 477 genes with no polymorphism and 16 with no divergence were removed from the analysis.

Calculation of diversity and selection statistics for 1-1 ortholog alignments between *H. melpomene* and *H. erato*: *Classic Approach*.

For the *H. melpomene* sequences orthologous with *H. erato* I calculated: 1) synonymous polymorphism (π_S) and 2) non-synonymous polymorphism (π_n). I also calculated synonymous divergence (dS), non-synonymous divergence (dN) between the *H. melpomene* and the *H. erato* sequences to estimate the rate of adaptive molecular evolution (alpha, α) between the two species. These values were calculated using the EggLib C++ function polymorphismBPP (v2.1.11) (De Mita and Siol, 2012) and Bio++ third-party library (v2.2.0) (Dutheil and Boussau, 2008) in python (v2.7.5) using scripts adapted from https://github.com/tatumdmortimer (O'Neill *et al.*, 2015).

Calculation of diversity and selection statistics for 1-1 ortholog alignments between *H. melpomene* and *H. erato*: *Modelling Approach*.

The *Modeling approach* estimates the strength of positive and purifying selection using the method of Eyre-Walker and Keightley (2009) as it was implemented in Galtier (2016) and Rousselle *et al.* (2016). The *Modeling approach* elaborates on the McDonald-Kreitman test by modeling the distribution of the fitness effect (DFE) of deleterious non-synonymous mutations as a negative Gamma distribution. The model is fitted to the synonymous and non-synonymous site frequency spectra (SFS) and the expected dN/dS under near-neutrality is inferred. The difference between the observed and expected dN/dS provides an estimate of the proportion of adaptive non-synonymous substitutions (α). The per mutation rate of adaptive substitutions is calculated as $\omega_a = \alpha(dN/dS)$; and the per mutation rate of non-adaptive substitutions is calculated as $\omega_{na} = (1 - \alpha)(dN/dS)$.

Results

Hmel2.1 annotation and 1-1 ortholog prediction with *H. erato*

There are 20 118 gene models predicted in *H. erato* (van Belleghem et al., 2017) and 13 019 predicted gene models in the *H. melpomene* Hmel2 released annotation (Davey *et al.*, 2016). These two annotations have 9 320 orthogroups, 6 846 of which are single-copy in the two species and 15 (0.2%) species-specific. 13 744 (68.3%) genes were assigned to an orthogroup in *H. erato* and 10 530 (80.9%) were assigned to an orthogroup in *H. melpomene*. Not all genes were assigned an orthogroup and most orthogroups are not single-copy orthogroups. This means more than one gene from each species can belong to an orthogroup (excluding single-copy orthogroups).

The *H. melpomene* Hmel2.1 updated annotation described here has 21 611 predicted gene models. When the same analysis is done between the Hmel2.1 release and *H. erato* the total number of orthogroups increases to 11 062; 8 085 of which are single-copy in the two species and 18 (0.3%) species-specific. 14 841 (73.8%) of genes were assigned to an orthogroup in *H. erato* and 14 857 (68.6%) were assigned to an orthogroup in *H. melpomene* (Figure 1A and 1B, Supplementary Table S2).

The updated gene set for H. melpomene is therefore much more comparable to the published *H. erato* gene annotation and is therefore more appropriate for future transcriptomic analysis in *H. melpomene*. We have made the annotation publicly available on LepBase (Challis *et al.*, BioRxiv preprint; also in https://www.dropbox.com/sh/5krc7kn3u0oviwj/AADHTlQsoxQCnqZnivatNdRba?dl=0).

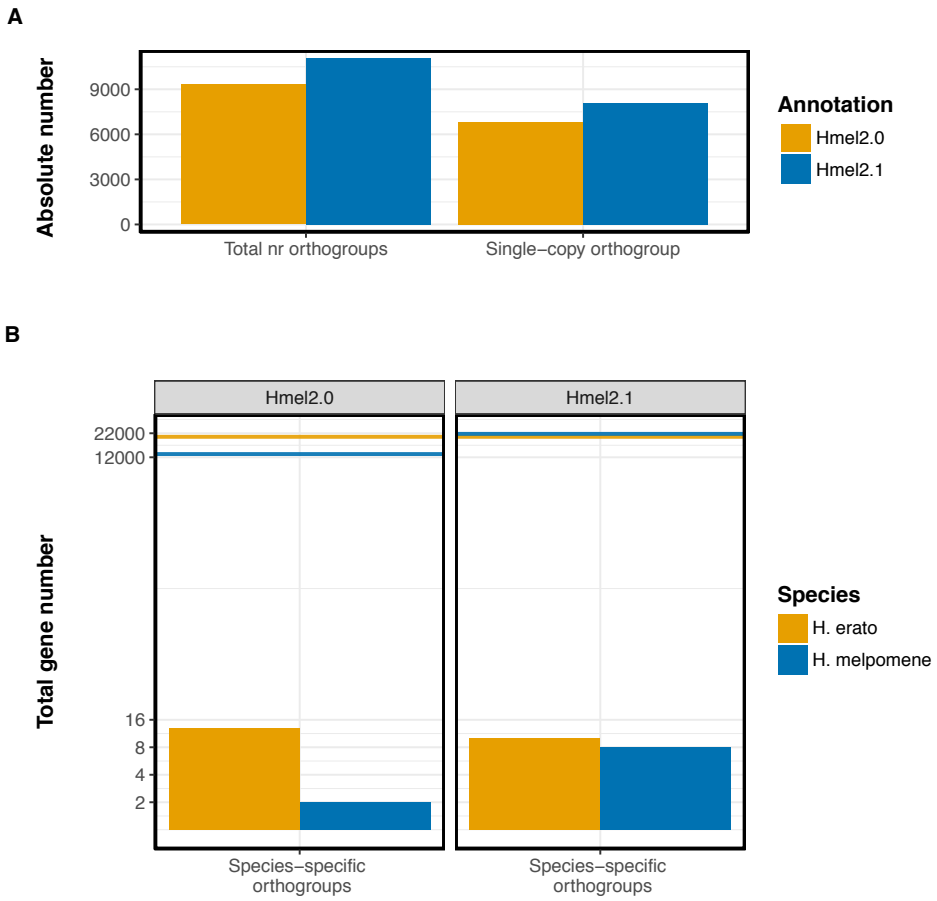


Figure 1. Hmel2.1 annotation yields comparable results to *H. erato* and it is more accurate than Hmel2.0

A. The absolute number of orthogroups and single-copy genes identified with OrthoFinder is larger in the Hmel2.1 annotation compared to Hmel2.0. **B.** The total number of putative genes is lower in the Hmel2.0 annotation compared to Hmel2.1, with the latter more comparable to the *H. erato* annotation. The number of species-specific orthogroups increases for *H. melpomene* when the Hmel2.1 annotation is used in the analysis against the *H. erato* published annotation.

RNAseq and read mapping

The 25 *H. melpomene* samples sequenced for this project have a median total number of reads of 34.86 M (min. 27.81 M; max. 46.12 M). The median total number of reads is similar to previously published gene expression studies in *Heliconius* (Briscoe *et al.*, 2013; Walters *et al.*, 2015). However, the experimental and sequencing protocol has greatly increased the number of both mapped reads and properly paired reads; and mapping success is high compared to other published studies (e.g. Yu *et al.*, 2016 and Walters *et al.*, 2015). Mapping success in the gut is lower than for the ovaries and the main driver of this is likely to be the gut microbiome.

For the gut samples the median percentage of mapped reads is 77.47%. For the ovary samples this value increases to 85.7%. Of the total number of sequenced reads, in the gut samples the median of the properly mapped reads is of 72.1%. For the ovary reads this value increases to 80.81%. These mapping statistics are a pronounced improvement over previously published data for *Heliconius*. For example, when I analysed the abdomen samples from Walters *et al.* (2015), there is a median 49.58% of mapped reads and only 43.97% of the total are properly paired. These mapping statistics are consistent to those originally reported for these data sets (Figure 2, Supplementary Table S1).

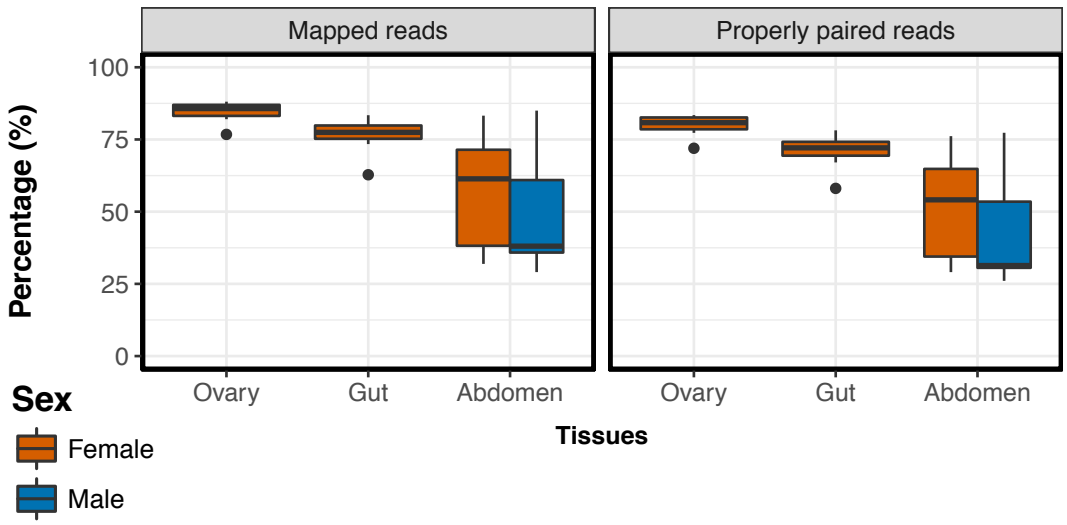


Figure 2. Percentage of mapped reads and properly paired reads of the samples used in this study

Box-and-whisker plot reporting the summary mapping stats of the samples sequenced for this study (Tissues: Ovary and Gut) and the samples downloaded from NCBI (Tissue: Abdomen).

Gene expression in whole-abdomen clusters individuals by sex

There is a clear separation of the 10 whole abdomen samples by sex when we compare gene expression profiles between them. In total, 98% of the total variance is explained by the two first principal components. PC1 separates the samples by sex and explains 97% of variance (Figure 3).

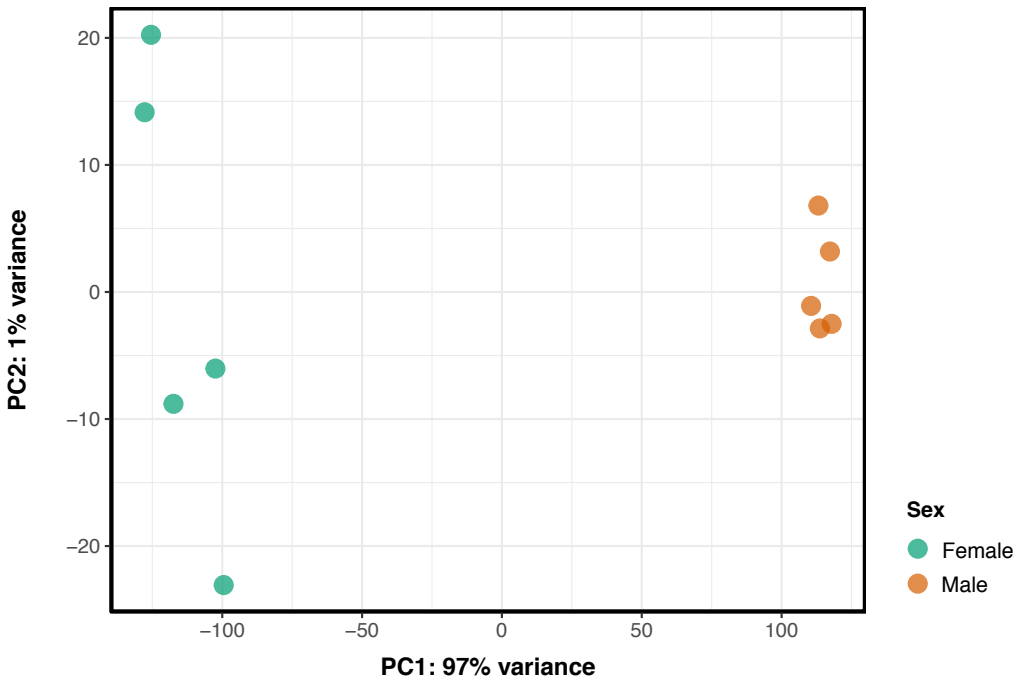


Figure 3. Principal component analysis of gene expression profiles for the 10 whole abdomen male and female samples

PCA of the abdomen transformed gene expression count data to the log2 scale (DESeq2, rlog(blind=FALSE)). rlog transformed data minimises differences between samples for rows with small counts and normalizes with respect to library size.

Mean expression level on the Z chromosome supports a mechanism for dosage compensation similar to eutherian mammals but not median expression levels

In the dataset I re-analysed from Walters et al (2015), the mean expression level in the Z chromosome is 60.55% lower than the autosomal mean expression level. Autosomal expression is similar in males and females, with female mean expression of autosomal linked genes 4.02% lower than male autosomal linked expression. Female mean expression of Z linked genes is, however, 29.75% lower than male mean expression (Figure 4). This difference is greater than what was reported using the previous annotation in the analysis of these data (Table 1).

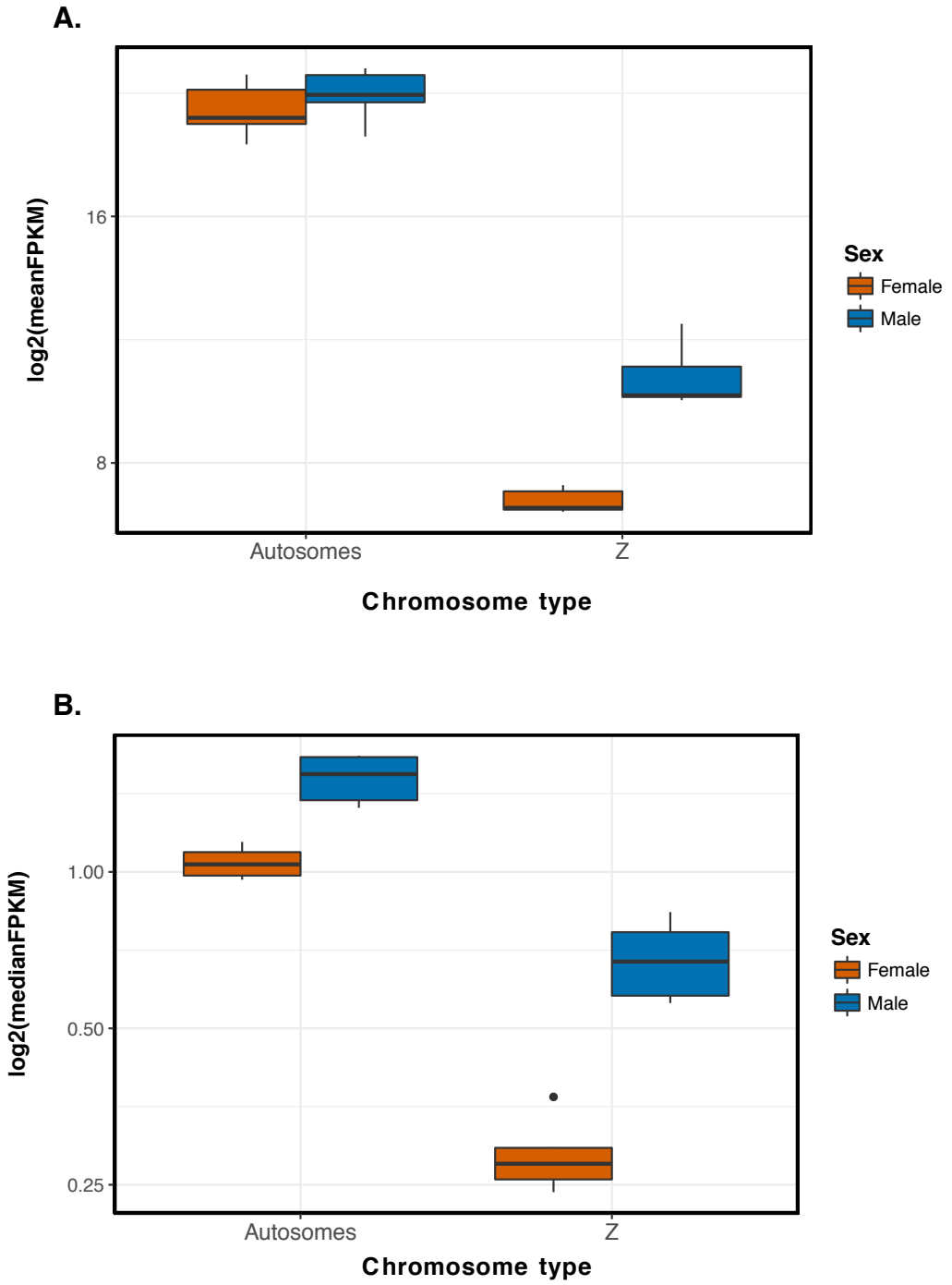


Figure 4. Distribution of mean and median expression level for Z-linked and autosomal-linked genes in male and female *H. melpomene* whole abdomen sample

Box-and-whisker plot reporting distribution of mean (**A**) and median (**B**) expression level for Z and autosomal-linked genes for whole abdomen samples downloaded from NCBI analysed with Hmel1.1 annotation.

	Hmel1.1 annotation	Hmel2.1 annotation
No. Z-linked loci	408	1 093
No. autosomal-linked loci	9 859	18 835
Mean Z log2(M:F)	0.2418	0.5093
Median Z log2(M:F)	0.1904	1.2726
Mean autosomal log2(M:F)	0.0106	0.0592
Median autosomal log2(M:F)	-0.0788	0.5329
Z:A ratio of mean	1.1738	0.3946
Z:A ratio of medians	1.2051	0.3606

Table 1. Mean and median M:F gene expression ratios for Z-linked and autosomal linked genes

Comparison between summary averages between M:F gene expression rations for Z-linked and autosomal-linked genes showing published results using Hmel1.1 annotation (Walters et al. 2015) and the results using Hmel2.1 annotation.

Z-linked and autosomal linked divergence does not support a significant fast-Z effect

dN/dS, computed by pairwise alignment for each 1-1 orthologous genes between *H. melpomene* and *H. erato*, has a slightly higher mean for Z-linked

genes than autosomal-linked genes. However, this is not significant and so there is not an obvious faster-Z divergence (Figure 5).

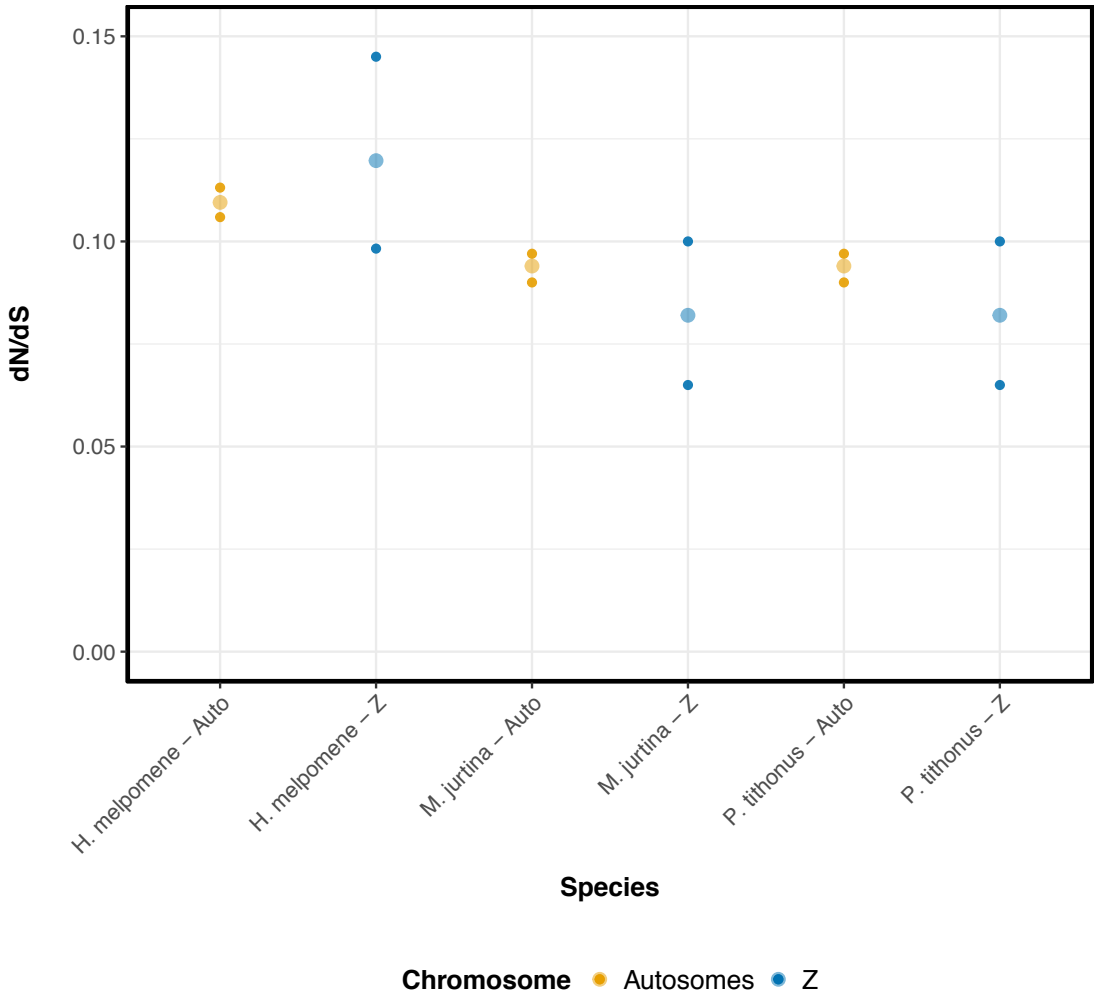


Figure 5. dN/dS

Values of dN/dS for autosomal and Z-linked genes for *H. melpomene* computed by pairwise alignment for each 1-1 orthologous genes between *H. melpomene* and *H. erato*; and the same values previously calculated for two satyrine butterflies (Rousselle *et al.*, 2016). Larger transparent points represent the median values of dN/dS. Smaller darker points, the upper and lower confidence intervals (1 000

replicates without replacement). dN/dS values for all genes irrespective of expression profile.

Z- and autosomal-linked polymorphism doesn't support reduced efficacy of purifying selection in Z-linked genes

I compared the levels of non-synonymous (π_n) and synonymous (π_s) polymorphism by calculating the π_n/π_s ratio between Z and autosomes (Supplementary Table S5). The Z chromosome π_s is lower than π_s for the autosomes. The average π_n/π_s ratio for Z-linked genes is slightly higher than autosomal-linked genes. The difference between π_n/π_s ratio of the Z and autosomes is, however, not significant (1 000 replicates without replacement) (Figure 6).

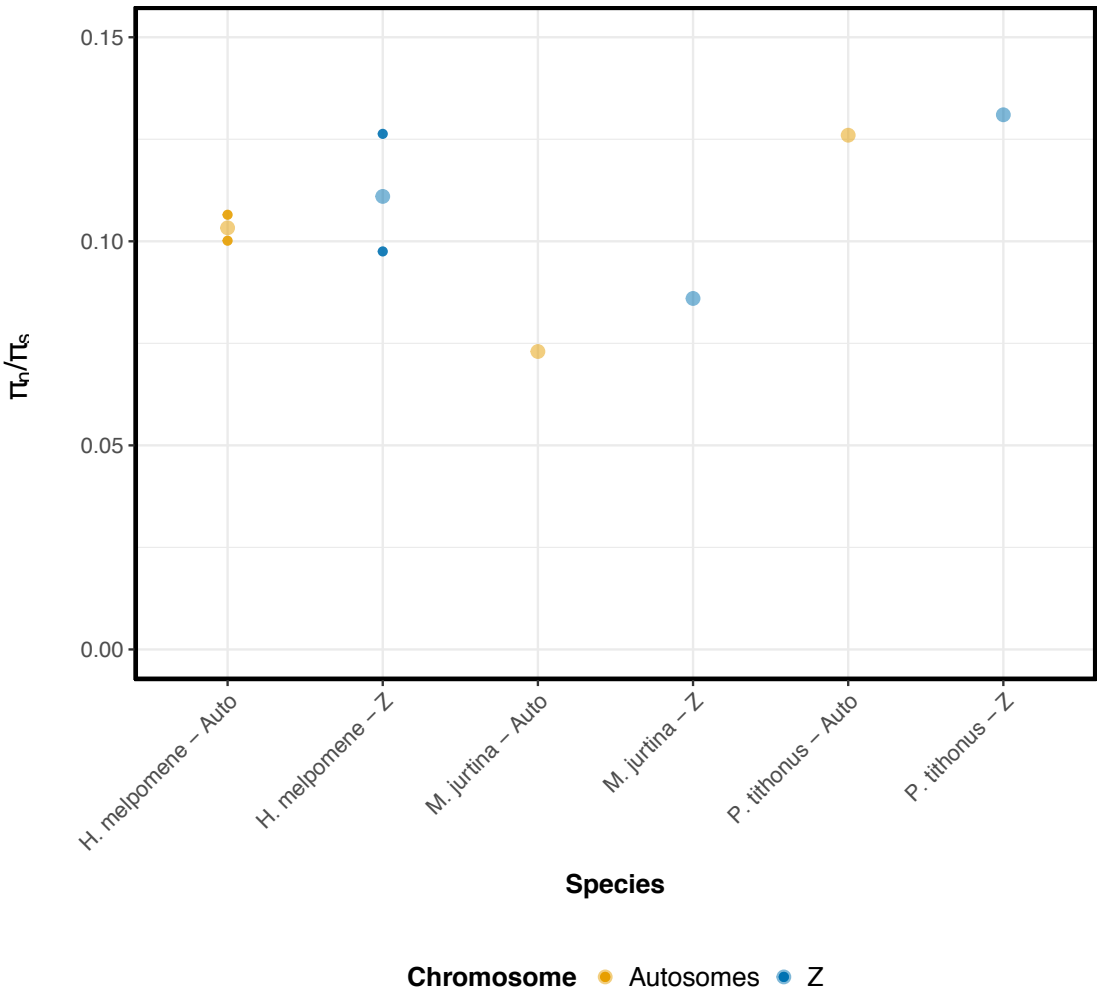


Figure 6. π_n/π_s

Values of π_n/π_s for autosomal and Z-linked genes for *H. melpomene* computed by pairwise alignment for each 1-1 orthologous genes between *H. melpomene* and *H. erato*; and the same values previously calculated for two satyrine butterflies. Larger transparent points represent the median values of π_n/π_s. Smaller darker points, the upper and lower confidence intervals (1 000 replicates without replacement). π_n/π_s values for all genes irrespective of expression profile.

Using the π_{sZ}/π_{sA} ratio to estimate Ne_Z/Ne_A, I can postulate that Ne of the Z chromosome is approximately 0.44 when all the expressed genes are considered. Female-biased Z-linked genes have a slightly higher Ne (0.64); and male-biased and unbiased Z-linked genes have an Ne that is more similar to the overall Z-linked gene Ne average (~0.42).

Purifying selection and sex-biased gene expression: Z-linked female-biased genes have the lowest π_n/π_s

π_n/π_s ratio was higher for male-biased genes than for female-biased and unbiased genes. For autosomal-linked genes unbiased expressed genes had lower π_n/π_s ratios than female-biased. For Z-linked genes female-biased genes have the lowest π_n/π_s but the difference was not statistically significant (Figure 7). Female biased gene counts were lower than the 2 083 previously reported in Walters *et al.* (2015). Male biased gene counts were equivalent to that previously reported in Walters *et al*. (2015), which totalled 1720. There were a total of 4 932 genes with unbiased expression as compared to 7 178 identified previously (Walters *et al*., 2015) (Table 2).

Sex-expression	Autosomes	Z
All	7464	200
Female	1231	28
Male	1238	96
Unbiased	4739	193

Table 2. Number of genes with sex-biased expression

Number of genes with female-biased, male-biased, unbiased expression for Z-linked and autosomal linked genes. Total number of expressed genes also shown (i.e. All). Sex-biased gene expression total gene number for Z-linked and autosomal-linked genes.

π_n/π_s of male-biased Z-linked genes is higher than Z chromosome and autosomal average. A higher π_n/π_s of male-biased Z-linked genes might indicate increased effect of genetic drift in the Z relative to the autosomes. Increased strength of genetic drift would promote segregation of slightly deleterious alleles in Z-linked male biased genes because they are lowly expressed or not expressed at all in females (heterogametic sex) (Figure 7).

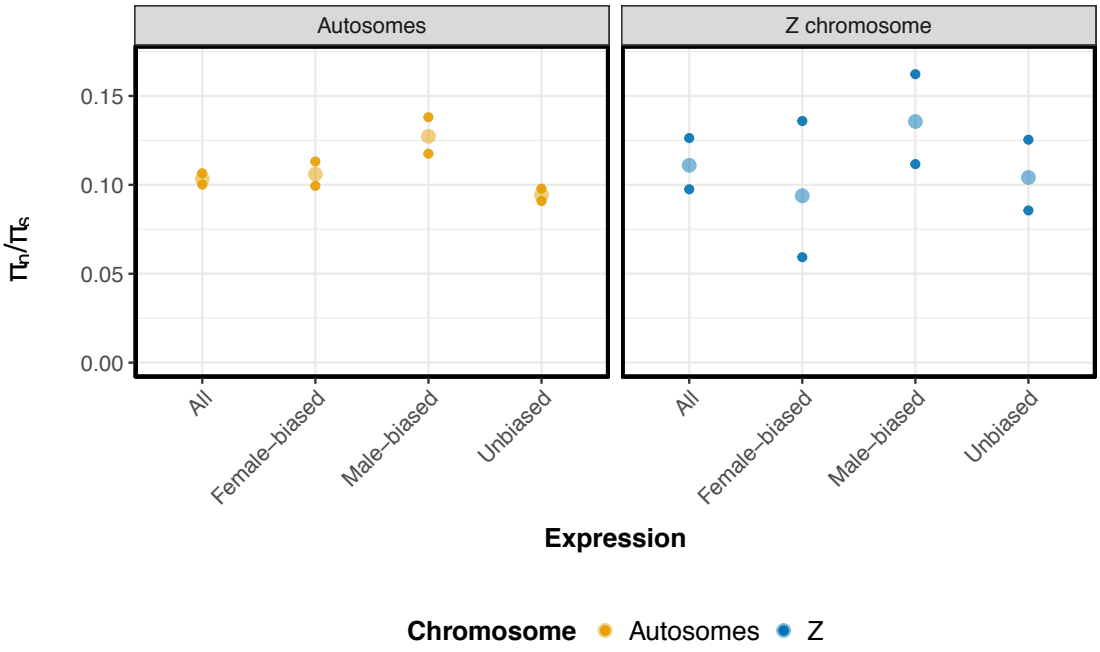


Figure 7. π_n/π_s for genes with sex-biased expression

Values of π_n/π_s for autosomal and Z-linked genes for *H. melpomene* computed by pairwise alignment for each 1-1 orthologous genes

between *H. melpomene* and *H. erato*. π_n/π_s ratios are split by sex-biased expression patterns. Larger transparent points represent the median values of π_n/π_s. Smaller darker points, the upper and lower confidence intervals (1 000 replicates without replacement).

Z linked genes have a median expression level significantly smaller than autosomal linked genes

Z linked genes have a median expression level significantly smaller than autosomal linked genes (Figure 8). Using a multiple regression approach we found that π_n and dN were significantly negatively correlated to expression level. This is true for π_n for both autosomal linked ($P < 0.001$) and Z linked genes ($P < 0.01$) and can be interpreted as increased strength of purifying selection on highly expressed genes (Figure 9, Table 3). The lack of a Z chromosome effect on π_n/π_s despite reduced expression and smaller effective population size means that there is no indication that the Z chromosome experiences reduced efficacy of purifying selection. However, this pattern was not observed for dN in the autosomes, which would in turn suggest an effect of hemizygosity on the efficacy of purifying selection (Table 4). The diagnostic plots for the linear regression analysis of gene expression and dNdS illustrates, however, that the model is a bad fit to the data and so there are likely to be non-linear relationships in the data that are not being captured by the model (Figure 10). For the analysis of each chromosome separately read *Supplementary Methods and Results* (Supp. Methods and Results Figure SM1, SM2 and SM3).

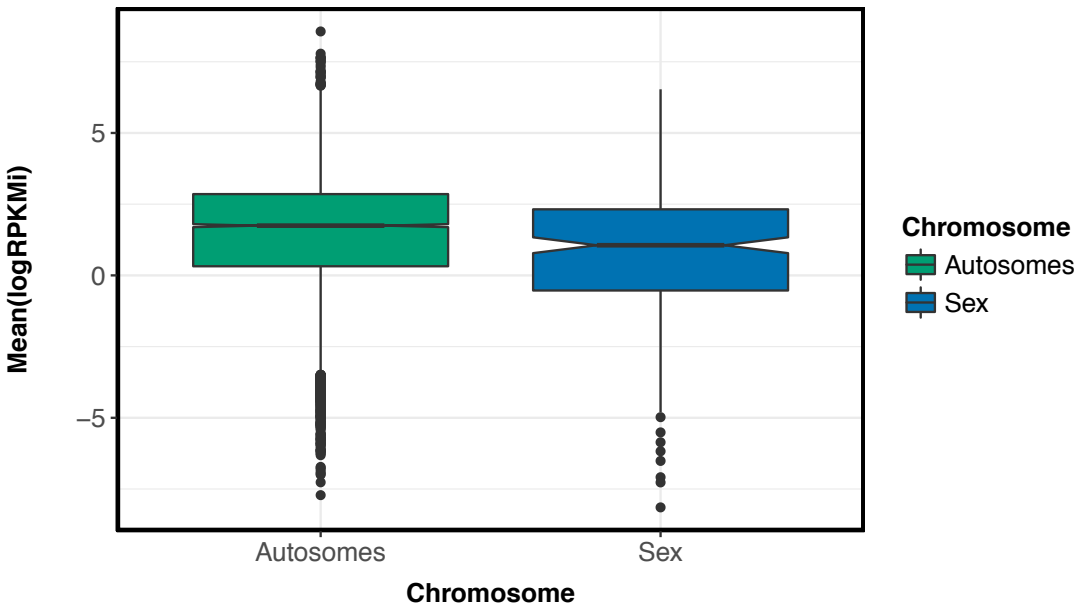


Figure 8. Median expression level of Z and autosomal linked genes

Median expression level of Z linked genes is significantly smaller than autosomal linked genes ($P < 0.05$). Notches on boxplot display the confidence intervals around the median.

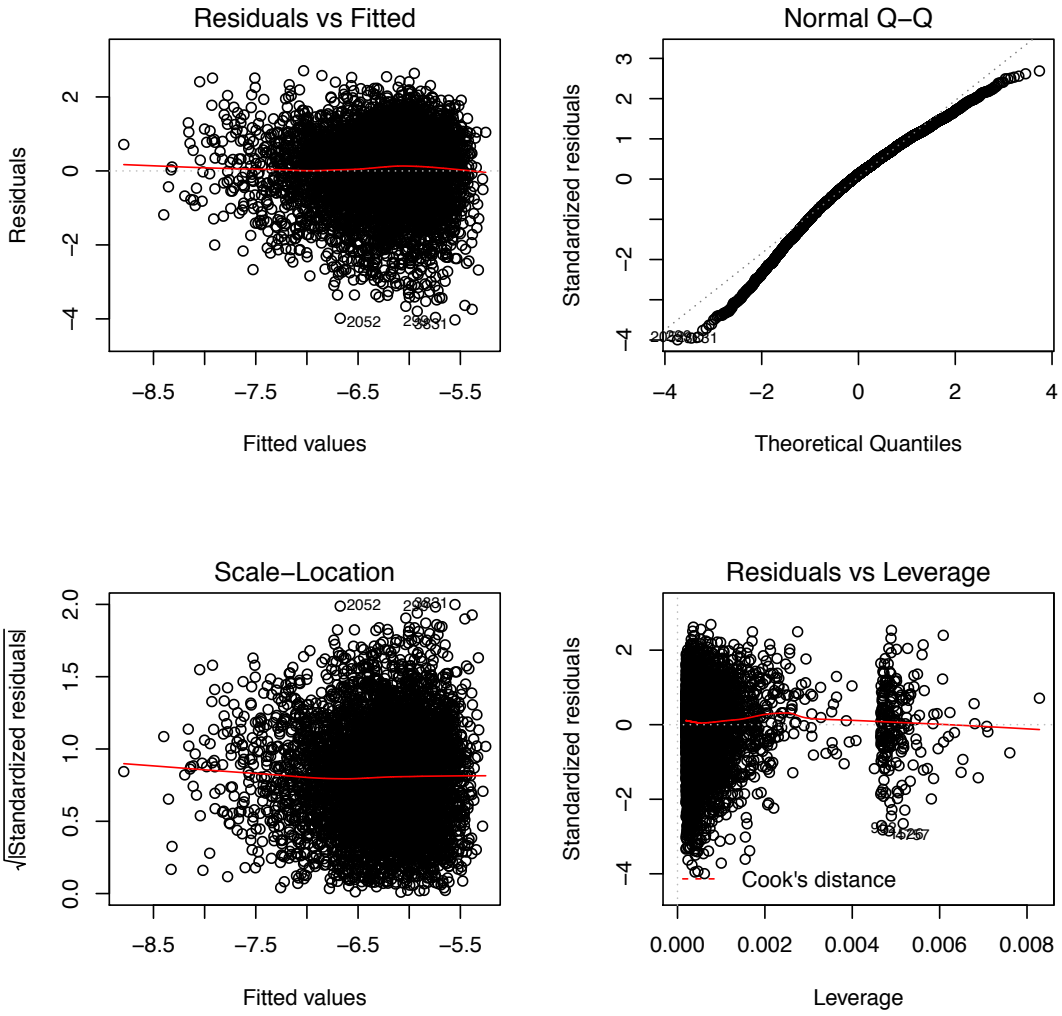


Figure 9. π_n is negatively correlated to expression level

Multiple regression approach shows that π_n was significantly negatively correlated to expression level – autosomal linked ($P < 0.001$) and Z linked genes ($P < 0.01$). Plotted *Residuals vs Fitted* shows spread residuals around the horizontal line without distinct patterns. *Normal Q-Q* follow a straight line with residuals well lined. The *Scale-Location* plot shows residuals spread equally around range of predictors. There is equal variance or homoscedasticity. *Residuals vs Leverage* plot does not identify any influential outliers in the linear regression analysis.

	Estimate	Std. Error	t value	Pr(>ltl)
(Intercept)	-4.318	0.060	-71.617	< 2e-16 ***
log(π_s)	0.469	0.015	31.164	< 2e-16 ***
chromosome_sex	-0.226	0.071	-3.174	0.002 **
log(RPKMi)	-0.035	0.006	-5.529	3.36e-08 ***

Table 3. Adjustment of the linear model $\log(\pi_{nij}) \sim \log(\pi_{sij}) + \text{chromosome\_type}_j + \log(\text{FPKM}_i)$

A multiple regression analysis was used to establish the following linear model $\log(\pi_{nij}) \sim \log(\pi_{sij}) + \text{chromosome\_type}_j + \log(\text{FPKM}_i)$ using R (v3.2.5). FPKM_i is the mean FPKM of gene *i* across the 10 individuals. 477 genes with no polymorphism were removed from the analysis. Multiple R-squared: 0.183, Adjusted R-squared: 0.182. F-statistic: 404.9 on 3 and 5428 DF, p-value: < 2.2e-16.

	Estimate	Std. Error	t value	Pr(>ltl)
(Intercept)	-3.242029	0.111055	-29.193	< 2e-16 ***
log(dS)	1.451338	0.031094	46.676	< 2e-16 ***
chromosome_sex	-0.002282	0.186207	-0.012	0.99
log(RPKMi)	-0.077614	0.016664	-4.658	3.25e-06 ***

Table 4. Adjustment of the linear model $\log(d_{Nij}) \sim \log(d_{sij}) + \text{chromosome\_type}_j + \log(\text{FPKM}_i)$

A multiple regression analysis was used to establish the following linear model $\log(d_{Nij}) \sim \log(d_{sij}) + \text{chromosome\_type}_j + \log(\text{FPKM}_i)$ using R (v3.2.5). FPKM_i is the mean FPKM of gene *i* across the 10 individuals. 16 genes with no divergence between *H. melpomene* and

H. erato were removed from the analysis. Multiple R-squared: 0.23, Adjusted R-squared: 0.30. F-statistic: 746.4 on 3 and 7497 DF, p-value: < 2.2e-16.

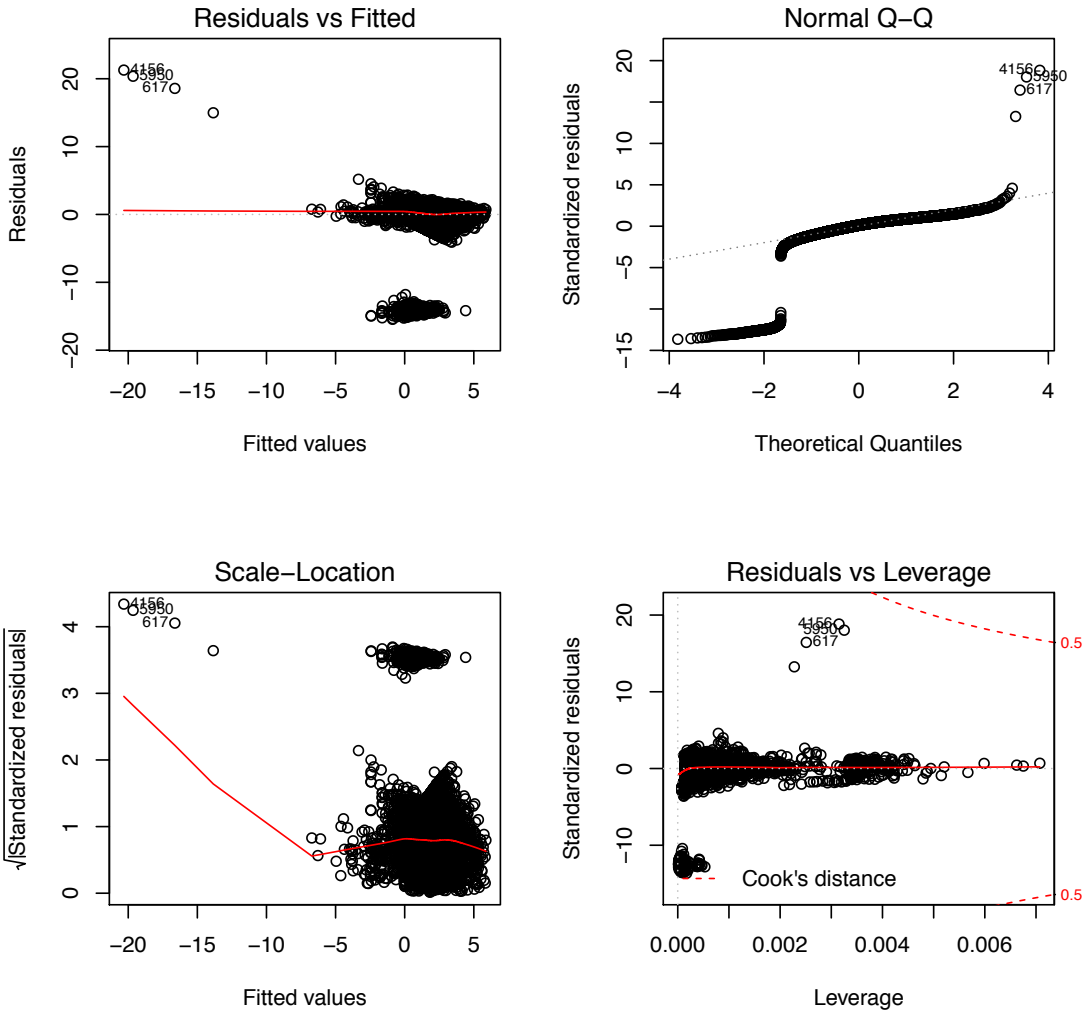


Figure 10. dNdS and expression level the model is a bad fit to the data

There is no equal spread of the residuals around the horizontal line so there may be non-linear relationships in the data The *Residuals vs Fitted* plot shows spread residuals around the horizontal line with distinct patterns. *Normal Q-Q* do not follow a straight line. The *Scale-Location* plot shows residuals unequally spread around range of

predictors. *Residuals vs Leverage* plot identifies influential outliers in the linear regression analysis that, even after being removed, do not improve the fit of the model significantly (not shown).

Z and autosomal rates of adaptive substitution: *Classic* and *Modelling Approaches* to test faster-Z adaptation

I first calculated rates of adaptive substitution (α) for each gene with a 1-1 ortholog between *H. erato* and *H. melpomene*. I will refer to such calculations of α as *Classic Approach*. To explore rates of adaptive substitutions and positive selection further I assessed the prevalence of adaptive evolution following a *Modelling Approach* where the effect of deleterious mutations is accounted for. This approach was first described by Eyre-Walker and Keightley (2009) and developed by Galtier (2009). Using this *Modelling Approach* we computed the proportion of adaptive non-synonymous substitutions α, as well as ω_a and ω_{na}. ω_a is the per mutation rate of adaptive substitutions and ω_{na} is the per mutation rate of non-adaptive substitutions.

Classic Approach: α is not significantly different between Z linked and autosomal genes

α ranges from 0 to 1 for autosomal and sex linked genes. α is has a mean value of 0.64 (median 0.71) for autosomal genes; and a mean value of 0.68 (median 0.75) for Z linked genes (Figure 11).

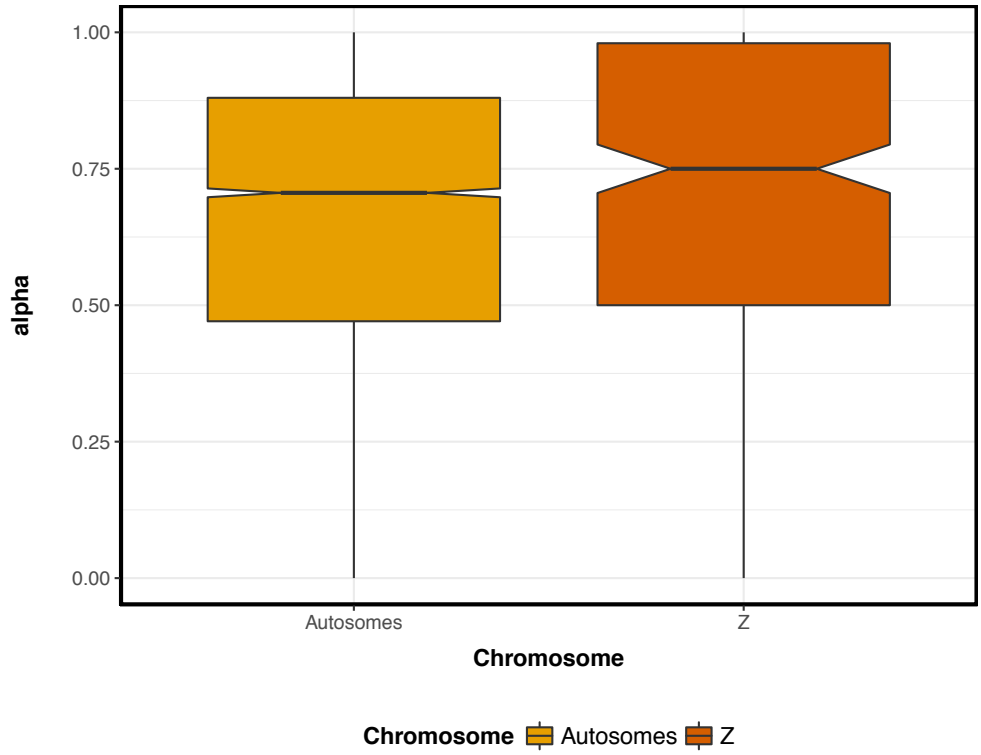


Figure 11. *Classic Approach*: Distribution of α in Z and autosomal linked genes

Box-and-whisker plot of α for autosomal and Z linked genes. α was calculated for each gene with a 1-1 ortholog between *H. erato* and *H. melpomene*. pS, pN, dN and dS were calculated between the reference *H. erato* CDS and CDS sequences extracted from 10 wild *H. m. rosina* samples.

Classic Approach: α is higher for genes with female biased expression patterns for Z linked genes than male biased but lower than unbiased Z linked genes

The median of α, irrespective of genomic location, is 0.70 for genes with female biased expression, 0.60 for genes with male biased expression and 0.73 for genes with unbiased expression between males and females (Figure 12, Table 5, *Classic Approach*).

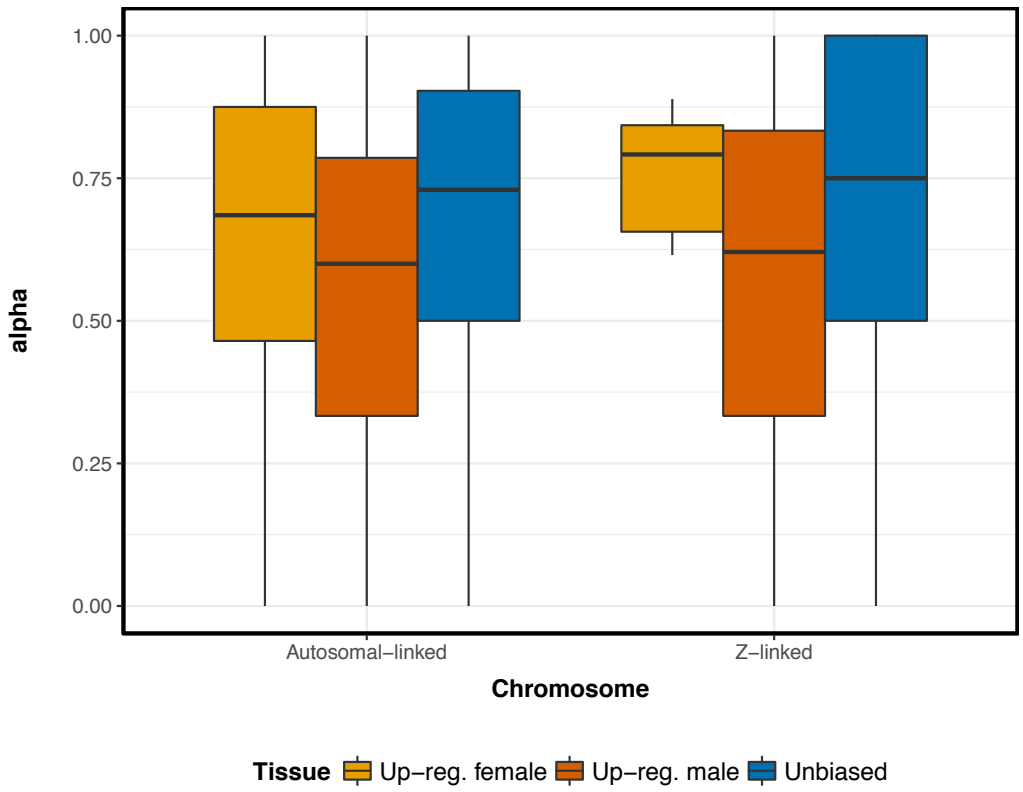


Figure 12. *Classic Approach*: Distribution of α genes with female biased, male biased and unbiased gene expression patterns accounting for genomic location

Box-and-whisker plot of α distributions plotted for each category on the y-axis.

	Autosomal-linked α	Z-linked α	Total α
Female biased	0.69	0.80	0.70

Male biased	0.60	0.62	0.60
Unbiased	0.73	0.75	0.73

Table 5. *Classic Approach*: Median α values for female biased, male biased and unbiased genes

Median α values calculated for each gene with a 1-1 ortholog between *H. erato* and *H. melpomene*. Genes are considered to exhibit biased expression patterns when they have a minimum fold-change of 1.5 and FDR<0.05 between males and females. Unbiased genes are equally expressed in females and males.

By bootstrapping a vector with all the values of α 5 000 times and calculating 95% confidence intervals for the regression coefficients (0.0116; 0.0261) accounting for genomic location, female biased Z-linked genes have a median value of α significantly higher than male biased genes but not higher than unbiased genes (std. error 0.038) (Figure 13).

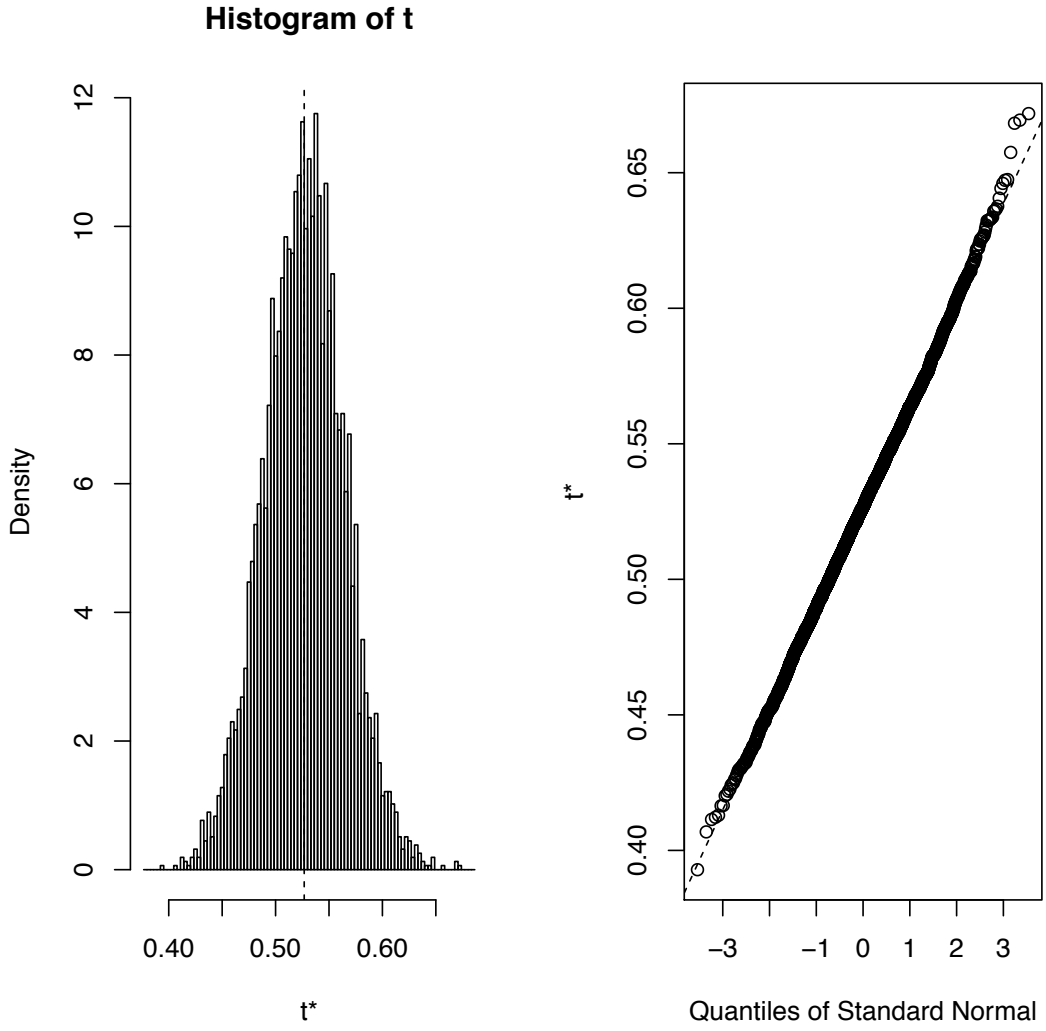


Figure 13. 5 000 nonparametric bootstrapping of the α vector

Random indices generated, with replacement, from the integers 1:8085. t is a matrix where each row is a bootstrap replicate of the statistics. 95% confidence intervals for the regression coefficients 0.0116; 0.0261. t* bias -3.103553e^{-05}; std. error 0.038.

Median α values calculated for each gene with a 1-1 ortholog between *H. erato* and *H. melpomene* indicate that positive selection is stronger on Z-linked female-biased genes than male-biased genes. However, median α

values are not significantly higher for Z-linked genes. To sum up, so far, using this *Classic Approach*, I have detected that positive selection seems to be slightly higher in Z-linked female-biased genes but the overall adaptive substitution rate is not higher in the Z chromosome compared to autosomes (i.e. there is no evidence of faster-Z adaptation).

Z-linked and autosomal-linked rates of adaptive substitution: results from the *Classic* and *Modeling Approaches*

Accounting for the effect of deleterious mutations by calculating mutation site-frequency spectrum we estimated the expected distribution of α, ω_a and ω_{na} (Table 6, *Modeling approach*).

	All	Female biased	Male biased	Unbiased
Autosomal-linked α	0.629 [0.622-0.636]	0.635 [0.620-0.650]	0.630 [0.616-0.646]	0.538 [0.529-0.547]
Z-linked α	0.675 [0.647-0.704]	0.699 [0.595-0.811]	0.646 [0.596-0.697]	0.537 [0.500-0.576]
Autosomal-linked ω_α	0.062 [0.061-0.063]	0.066 [0.065-0.068]	0.087 [0.085-0.089]	0.047 [0.046-0.048]
Z-linked ω_α	0.069 [0.072-0.066]	0.069 [0.058-0.080]	0.090 [0.083-0.097]	0.048 [0.044-0.051]
Autosomal-linked ω_{na}	0.036 [0.036-0.037]	0.038 [0.037-0.040]	0.051 [0.049-0.053]	0.040 [0.039-0.041]
Z-linked $\omega_{n\alpha}$	0.033 [0.030-0.036]	0.029 [0.019-0.040]	0.049 [0.042-0.056]	0.041 [0.038-0.044]

Table 6. *Modeling approach:* Estimated rates of adaptation (α), adaptive (ω_α) and non-adaptive substitution rates ($\omega_{n\alpha}$)

Values computed with the method of Eyre-Walker & Keightley (2009). Intervals represent 95% confidence intervals obtained by bootstrapping with 1 000 times.

By comparing median α values calculated for each gene with a 1-1 ortholog between *H. erato* and *H. melpomene* (i.e. *Classic Approach*) to those estimated with the Eyre-Walker & Keightley (2009) (i.e. *Modeling approach*) is we can see that the *Classic Approach* generally calculates higher median values of α. This is particularly the case for autosomal-linked and Z-linked genes with unbiased expression. *H. melpomene* male-biased autosomal and Z-linked have higher values of α for the *Modeling* approach but the confidence intervals overlap the *Classic Approach* values. As in the *Classical Approach*, in the *Modelling Approach* female-biased Z-linked genes have a higher α than male-biased and unbiased genes. However, the confidence intervals are broad and this is not significant suggesting further that hemizygosity does not have a strong effect on the rate of adaptive substitution. Overall α is similar between the Z chromosome and autosomes (Table 5 and 6, Figure 12 and 14).

It is somewhat surprising, however, that the α values of the *Classical approach* are higher than those of the *Modelling approach*. The *Classical approach* is expected to overestimate the proportion of non-adaptive divergence. Further work is needed in order to investigate this apparent discrepancy.

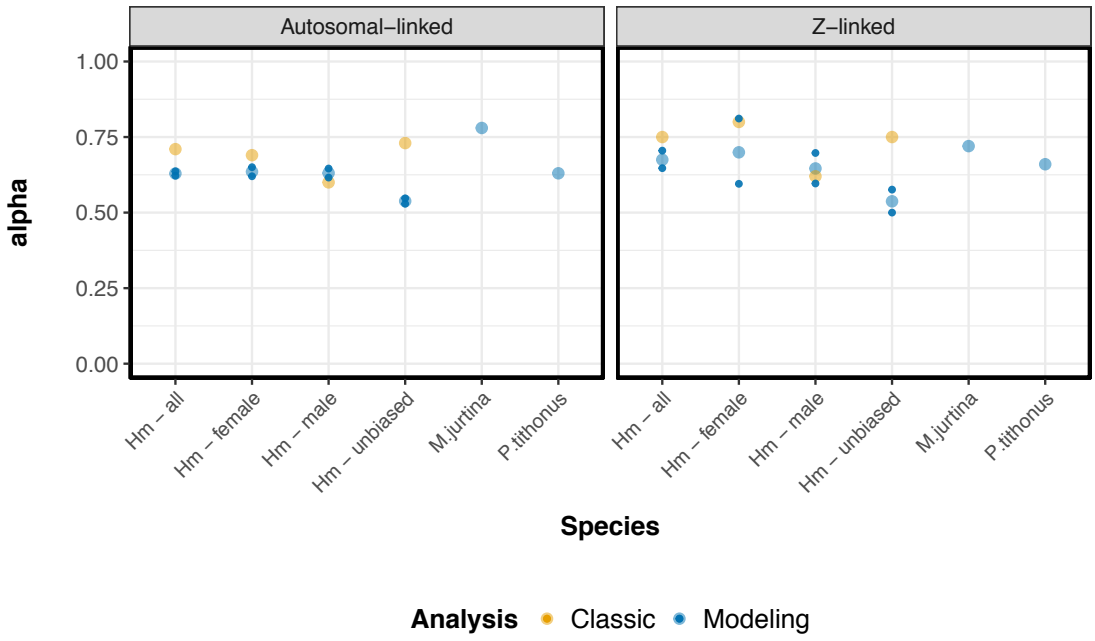


Figure 14. Values of α from the *Classic* and *Modeling Approaches*

Values of α from the *Classic* and *Modeling Approaches* for autosomal and Z-linked genes. α values for two satyrine butterflies irrespective of genomic location shown for comparison. Larger transparent points represent the median values of α. Smaller non-transparent points, the upper and lower confidence intervals for the *Modelling Approach* estimates of α. Hm, *Heliconius melpomene*. Hm – all, genes with 1-1 orthologues between *H. melpomene* and *H. erato* irrespective of expression profile. Female, *H. melpomene* female biased expression. Male, *H. melpomene* biased expression.

Hemizygosity might affect the rate of adaptive substitutions

ω_α is significantly higher for Z-linked genes compared to autosomal linked genes when all *H. melpomene* genes are considered. ω_α is also higher for female-biased, male-biased genes and unbiased Z-linked genes but the

difference between sex-chromosome and autosome is not significant. This might suggest that hemizygosity has an effect on the rate of adaptive substitution in *Heliconius* as the prevalence of positive selection in all Z-linked genes taken together is significantly higher than that of autosomes. There are no significant differences between autosomes and sex-chromosomes for $\omega_{n\alpha}$ (Supplementary Table S6). Female-biased genes have the lowest $\omega_{n\alpha}$ compared to male-biased and unbiased genes which confirms the low π_n/π_s already reported (Figure 15).

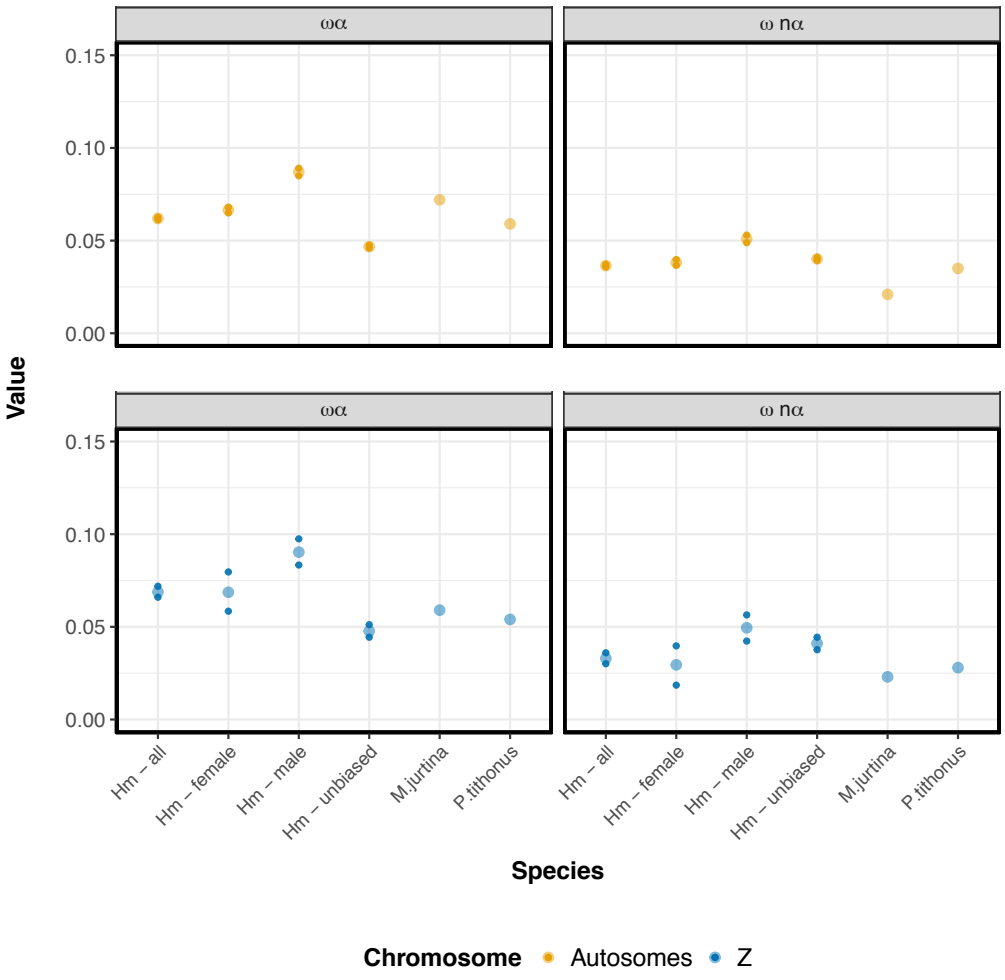


Figure 15. ω_α and $\omega_{n\alpha}$

ω_α and $\omega_{n\alpha}$ plotted for autosomal-linked genes and Z-linked genes. Larger transparent points represent the median values of α. Smaller

non-transparent points, the upper and lower confidence intervals for the *Modelling Approach* estimates of α. Hm, *Heliconius melpomene*. Hm – all, genes with 1-1 orthologues between *H. melpomene* and *H. erato* irrespective of expression profile. Female, *H. melpomene* female biased expression. Male, *H. melpomene* biased expression.

Proportionally there are more female-biased Z-linked genes than male-biased Z-linked genes. Z-linked female biased genes have the lowest π_n/π_s compared to both unbiased and male-biased genes. dS on the Z chromosome is higher than dS on the autosomes which is consistent with a male-biased mutation rate and not a strong fast-Z effect (Supplementary Table S4). $\omega_{n\alpha}$ is the lowest for female-biased genes; and dN/dS is also lower for female biased Z-linked genes than for autosomal genes. All of the above are consistent with a scenario in which hemizygosity in females affects the efficacy of purifying selection against recessive deleterious mutations. However, α is marginally higher for female-biased genes compared to male biased or unbiased genes. Moreover, ω_α is significantly higher for Z-linked genes overall and it is higher for female linked genes compared to unbiased genes (but lower than male-biased genes) (Supplementary Table S6). Sex-biased genes are disproportionally expressed in sex specific tissue like the testis and the ovaries. To dissect this further, and determine if genes expressed in female specific tissue (i.e. ovary) also showed the same molecular evolution patterns, I collected, extracted, sequenced and analysed ovary and gut tissue from young and old *H. melpomene* females.

Gene expression in female germline and somatic tissue clusters individuals by tissue and age

I analysed data from two different time points and from somatic and germline female tissue separately (Treatment: Young and Old). There is a clear

separation of the 25 samples by tissue when we compare gene expression profiles between them. In total, 51% of the total variance is explained by the two first principal components. PC1 separates the samples by tissue and explains 40% of variance. PC2 separates samples by age (Figure 16). Germline tissue (Ovary) clusters by age more tightly than somatic tissue (Gut) (Figure 16, Figure 17A and 17B).

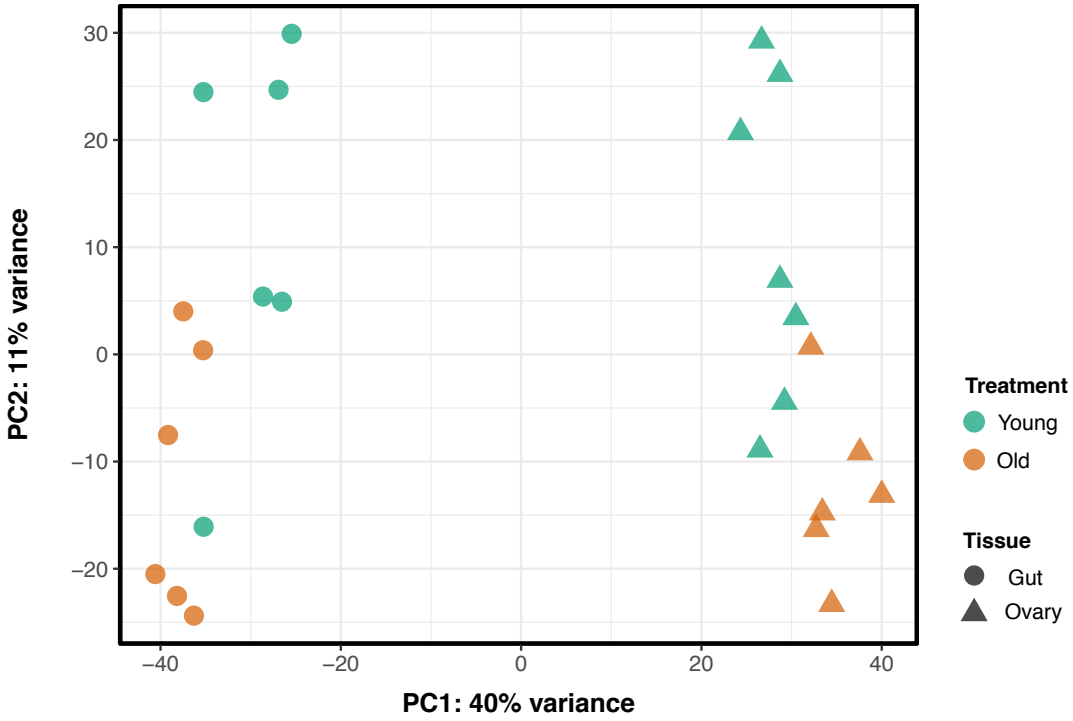


Figure 16. Principal component analysis of gene expression profiles of *H. melpomene* females for 13 ovary samples and 12 gut samples at two different time points

PCA of the female ovary and gut transformed gene expression count data to the log2 scale (DESeq2, rlog(blind=FALSE)). rlog transformed data minimises differences between samples for rows with small counts and normalizes with respect to library size.

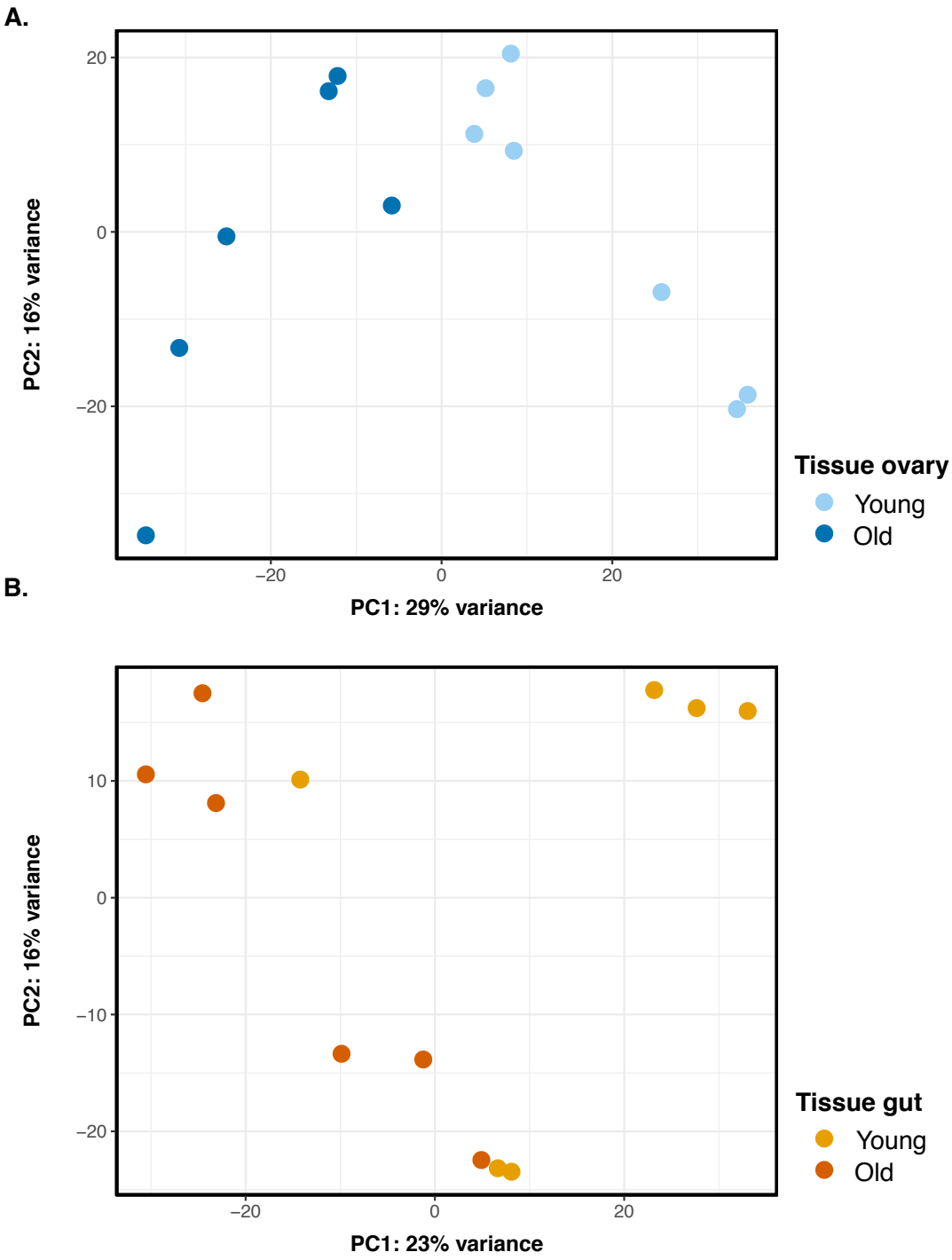


Figure 17. Principal component analysis of gene expression profiles of *H. melpomene* females for 13 ovary

samples and 12 gut samples at two different time points separated by tissue type

PCA of the female ovary and gut transformed gene expression count data to the log2 scale (DESeq2, rlog(blind=FALSE)) separated by tissue. rlog transformed data minimises differences between samples for rows with small counts and normalizes with respect to library size. **A.** 45% of the variance is explained by PC1 and PC2. PC1 separates young ovary tissue from old ovary tissue and explains 29% of the variance. All the samples cluster by age. **B.** 39% of the total variance is explained by PC1 and PC2. PC1 separates young gut tissue from old gut tissue and explains 23% of the variance. The samples cluster less tightly by age than ovary expression.

Overall there are a greater number of genes with gut-biased expression than ovary biased expression in the autosomes. However, there seems to be an over-representation of Z-linked ovary expressed genes. These results should be interpreted cautiously as the number of genes in each category is considerably smaller than in the whole abdomen samples (Table 7, Figure 18).

	Autosomal-linked	Z-linked	Total
Gut	153	6	159
Ovary	40	6	46

Table 7. Counts of sex-biased genes in gut and ovary tissue for autosomes and sex chromosome (Z)

Genes show biased expression patterns when they have a minimum fold-change of 1.5 and FDR<0.05. *Gut* encompasses genes that are up-regulated in the guts and *Ovary* in the ovaries.

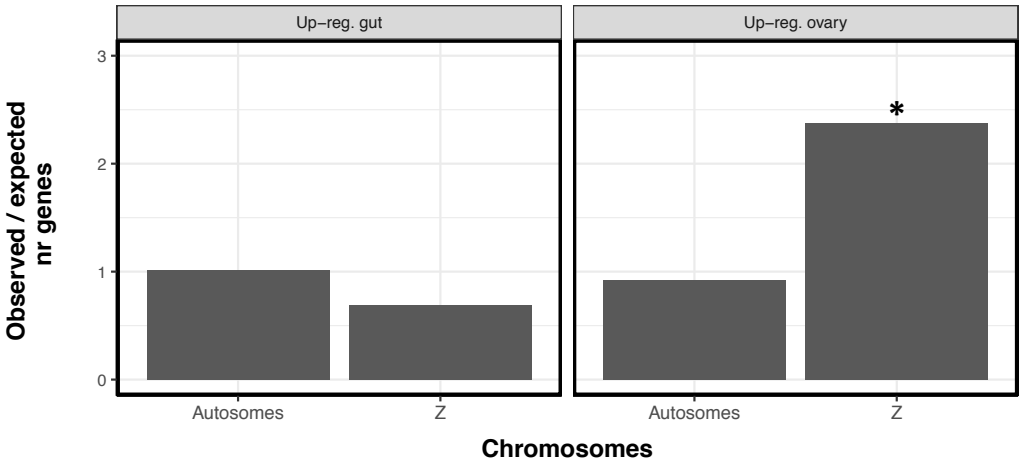


Figure 18. Female gut and ovary biased genes with autosomal or Z linkage

Proportion of observed/expected number of gut biased and ovary biased genes for female tissues. A value of 1 in the y-axis would indicate the same number of observed and expected genes. The asterisk indicates a significantly higher number of Z-linked genes expressed in male abdomen (chi-square test; p<0.05) than expected by chance.

Of the total 205 differentially expressed genes between the two tissues only 17 in the ovaries and 64 in the guts could be used to calculate dN/dS, piN/pS and α. The other genes either do not have a 1-1 ortholog with *H. erato* or there were too many undetermined characters (gaps or Ns) to be able to estimate the parameters. Of the 81 genes from which molecular evolution statistics could be calculated from, all ovary-biased (17 genes) and 63 gut-

biased are autosomal linked; and 1 gut-biased is sex-linked (HMEL008785, Hmel221009:209905-223861).

No significant differences in rate of adaptive evolution, positive selection or purifying selection between female ovary-biased and gut-biased genes

Median α values calculated for each one of the ovary-biased and gut-biased genes (autosomal and sex linked together) indicate that positive selection is stronger on ovary-biased genes but not significantly so. Mean dN/dS values are not significantly different for gut-biased genes versus ovary-biased genes. Even though the top quartile of gut biased genes dNdS > 1 there is no significant positive selection. The difference between π_n/π_s ratios between the ovary and gut-biased genes is also not significant (Figure 19).

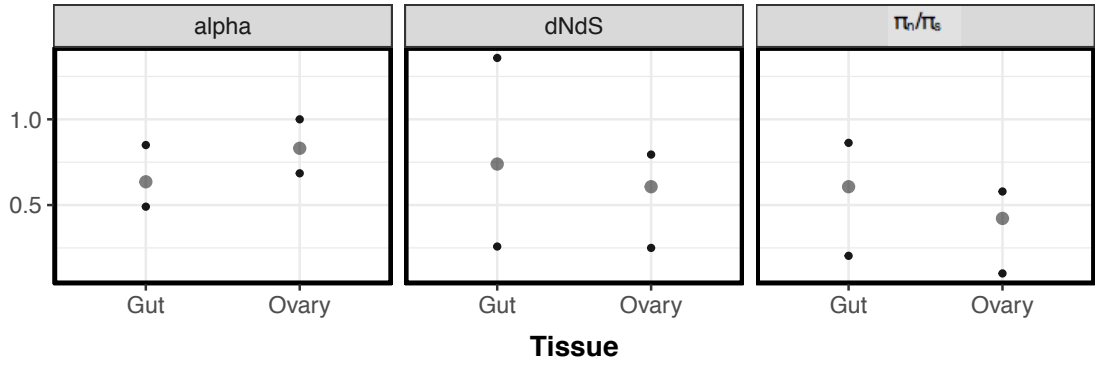


Figure 19. α, dN/dS and π_n/π_s for ovary and gut-biased genes

Values of α, dN/dS and π_n/π_s for gut and ovary-biased genes. Larger points represent the median values of α, dN/dS and π_n/π_s. Smaller

darker points, the upper and lower confidence intervals (1 000 replicates without replacement).

Discussion

I did not find evidence for fast-Z evolution in *H. melpomene*. Estimates of α, dN/dS and π_n/π_s were similar between Z-linked and autosomal-linked genes and there were no significant differences in the rate of adaptive evolution, positive selection or purifying selection between female- and male-biased genes or female ovary-biased and female gut-biased genes despite a low effective population size of the Z chromosome relative to autosomes.

Elevated rates of coding sequence evolution on the sex chromosome relative to autosomes have been reported for several species and, for taxa with complete dosage compensation mechanism, there is strong support for fast-X evolution (Mank, Vicoso, *et al.*, 2010; Meisel and Connallon, 2013). Specifically, opportunities for fast-X evolution are predicted to increase in species where there is somatic X-inactivation as it is observed in eutherian mammals. In taxa where there is somatic X-inactivation, within individual female cells, there is haploid expression which results in an increase of the chances of recessive beneficial mutations to be fixed (Charlesworth *et al.*, 1987).

Patterns of sex chromosome dosage compensation in female-heterogametic taxa are complex, but the reanalayis carried out here broadly confirms previous results in *Heliconius*. Previous work showed that *Heliconius* males have reduced the expression of Z-linked genes below autosomal expression, but this dosage compensation mechanism was imperfect with males showing increased expression relative to females on the Z chromosome (Walters *et al.*, 2015). Hence, there was reduced Z-linked expression relative to autosomes in both sexes but there was also a slight dosage effect on the Z chromosome (i.e. Z<ZZ<AA mechanism of dosage compensation). Using the updated *H.*

melpomene annotation, the median Z log2(M:F) is consistent with a mechanism of dosage compensation Z~ZZ<AA but the average is consistent with a Z<ZZ<AA pattern. A pattern of Z<ZZ~AA has been consistently reported in WZ systems with the exception of lepidoptera where Z<ZZ~AA; Z~ZZ<AA and Z<ZZ<AA have all been seen. In a recent study in *Cydia pomonella*, showed that in somatic tissue there is a dosage compensation mechanism similar to eutherian mammals (Z~ZZ<AA) but, in germline tissue, there is total absence of dosage compensation (Gu et al. 2017). The tissue I used to calculate the ratios has both somatic and germline tissue analysed together and so the average Z log2(M:F) might be lower than the median due to the fact that Z-linked germline expressed genes do not have a dosage compensation mechanism, but somatic ones do. It would therefore be interesting to analyse germline male expression against germline female tissue without somatic tissue contamination. Regardless, the lack of a complete dosage compensation mechanism in *Heliconius* is likely to contribute to the lack of fast-Z evolution in this species.

Although a fast-Z effect has been observed in *Bombyx mori*, most ZW systems analysed until now report the opposite (Sackton *et al.*, 2014). The lack of a fast-Z effect was also observed in two satyrine butterflies where dN/dS ratio of Z-linked genes was slightly lower than autosomal (Rousselle *et al.*, 2016). In *Heliconius*, similarly dN/dS is not significantly different between autosomal-linked and Z-linked genes. Moreover, dS on the Z chromosome is higher than dS on the autosomes perhaps indicating a male-biased mutation rate (Miyata *et al.*, 1987).

dNdS values may be upwardly biased at short time scales or in regions of low mutation rate (Mugal *et al.*, 2014). The *H. erato – H. melpomene* split is ~ 10.5 MYA (Kozak *et al.*, 2015 using BEAST) and Ne is roughly 2 to 3.5 million which would imply a split 3-5N generations ago. According to Mugal *et al.* (2015), and considering the amount of shared polymorphisms, dNdS values may be inflated and the data should be interpreted with caution. However, for this analysis, I compared dNdS and calculated the α for different classes of

genes (i.e. did not compare individual genes). This means that, even if the values are potentially inflated, all genes are equally affected and the comparisons among the difference genes categories are still valid.

In a fast-Z scenario, male-biased genes should not be affected by hemizygosity as the fast-Z is driven by recessive beneficial mutations. Wright *et al*. (2015) interpreted the high dN/dS in Z-linked genes of birds as a consequence of reduced effective population size rather than positive selection. The difference in effective population size between sex chromosomes and autosomes in female heterogametic systems is predicted to be larger than in male heterogametic systems due to higher variance of male reproductive success (Mank, Nam, *et al*., 2010). As in satyrine butterflies, in *Heliconius*, ω_{na} is not higher on the Z relative to autosomes. dN/dS and π_n/π_s are higher in the Z relative to autosomes in *Heliconius* however, this is not significant and I only detect reduced efficacy of purifying selection in male-biased genes. This means that, in contrast to birds, the difference in the effective population size of the Z relative to autosomes is not sufficient to reduce the efficacy of purifying selection at a detectable level. As hypothesised for the satyrine butterflies, and in contrast to what is observed in birds, an overall high effective population size could be one explanation of the differences between birds and lepidoptera (Rousselle *et al*., 2016).

Species with large effective population size are more polymorphic than small populations. Large amounts of polymorphism increase the probability of adaption from standing genetic variation, which in turn reduces, or completely eliminates the opportunity for fast-Z evolution (Charlesworth *et al*., 1987). Adaptation using standing genetic variation results in faster-autosome substitution, independent of the dominance of beneficial alleles as autosomes harbour more genetic diversity (Orr and Betancourt, 2001). *H. melpomene* has a large effective population size and so the differential use of *de novo* mutations versus standing genetic variation during adaptation may reduce the opportunities for fast-Z evolution.

Heliconius Z-linked female-biased genes have the lowest π_n/π_s and the lowest $\omega_{n\alpha}$ but this difference is not statistically significant. To summarise, the difference between π_n/π_s ratio of the Z and autosomes is not significant and so, despite low Ne of the Z chromosome relative to the autosomes there is no evidence for reduced efficacy of purifying selection. The effective population size of the Z chromosome is expected to be 0.75, ¾ of the autosomes. However, our data suggests that, in *Heliconius*, the effective population size of the Z is ~0.42. This value is smaller than the Ne of the Z calculated for *P. tithonus* (~0.6) but larger than the Ne of the Z calculated for *M. jurtina* (Rousselle *et al.*, 2016).

Departure from the expected Ne ratios between autosomes and sex chromosomes may result from difference process. For example, 1) sex-biased mutation rates (Miyata *et al.*, 1987); 2) sex-specific variance in reproductive success (Charlesworth 2001); 3) sex-biased migration (Laporte and Charlesworth 2002); 4) linked negative selection (Charlesworth 1996); 5) positive selection (Aquadro *et al.*, 1994); and 6) historical changes in population size, such as bottlenecks (Pool and Nielsen, 2007) can all contribute to reduce the Ne of the Z chromosome. In *Heliconius* several of these processes might be influencing the observed ratio. As previously discussed there may be a male-biased mutation rate. There is sex-specific variance in reproductive success, with males have a greater reproductive success variance than females; and there is sex-biased migration, with males migrating greater distances than females (Mallet 1986). However, as shown through my analysis, positive and negative selection on the Z may only affect the effective population size ratio mildly. Finally, demographic changes such as recent bottlenecks do not seem to correctly describe the demographic history of *Heliconius melpomene* and so it is less likely that this process has a considerable impact in the observed Z Ne reduction (Kozak *et al.*, 2015, Camille Roux pers. comm.). Regardless of the processes driving Ne reduction, if the efficacy of purifying selection was reduced in *Heliconius* due to low effective population size of the Z chromosome relative to the

autosomes I would expect higher π_n/π_s ratios on the Z compared to autosome due to stronger genetic drift. For example, high N_{ex}/N_{eA} ratios in *Drosophila* may explain why there is robust evidence for faster-X adaptation but not divergence (Meisel *et al.*, 2012).

As sex-biased genes tend to be expressed in sex specific tissue such as the testis and the ovaries I aimed to investigate patterns of molecular evolution in ovary-biased genes. Unfortunately, there are no ovary-biased genes with 1-1 orthologues between *H. melpomene* and *H. erato* that are Z-linked. This meant I could not test the effect of hemizygosity on somatic and germline female expression directly. The lack of 1-1 orthology between ovary-biased *H. melpomene* genes and *H. erato* may mean that these genes are under strong positive selection. However, I was still able to perform the analysis with autosomal linked genes. Ovary-linked genes have higher rate of adaptive evolution than gut genes, which is consistent with sex-linked genes having higher rates of adaptive evolution. In *Heliconius*, α values for autosomal and Z-linked genes are lower than have been estimated for *M. jurtina* but higher than estimates for *P. tithonus* (Rousselle *et al.*, 2016).

I identified sex-biased genes which include genes that are expressed 1) just in one sex (sex-specific expression) and, 2) in both sexes but at a higher level in one sex (sex-enriched expression). In the future, it would be interesting to disentangle the two and, perhaps, by coupling gene expression data with WGS data it might be possible to map W-linked regions. Specifically, through the identification of SNPs and genomic regions on female WGS data not present in the male *Heliconius*, we may identify female specific loci.

Together these results illustrate the need to study substitution rates in other ZW systems considering sex-biased expression. This genome-wide analysis of polymorphism, divergence and gene expression data contributes to a growing body of literature on sex chromosome evolution in ZW systems, and reveals the complexity of the different evolutionary forces shaping

transcriptome evolution in *Heliconius* and, consistent with previous work, shows the lack of fast-Z evolution in this taxon.

Supplementary Tables

Supplementary Table S1. Sample information and statistics

Sample	Sex	Tissue	Treatment	Library	Raw Reads
AP141	Female	Gut	Old	RRB03031	32306468
AP93	Female	Gut	Young	RRB03025	34353682
AP142	Female	Gut	Old	RRB03032	34445514
AP77	Female	Gut	Old	RRB03030	32024898
AP89	Female	Gut	Old	RRB03034	39661158
AP55	Female	Gut	Old	RRB03035	35862378
AP94	Female	Gut	Young	RRB03026	36223403
AP37	Female	Gut	Young	RRB03029	31187077
AP34	Female	Gut	Young	RRB03028	35452206
AP71	Female	Gut	Young	RRB03024	33088963
AP35	Female	Gut	Young	RRB03027	46122142
AP80	Female	Gut	Old	RRB03033	40040750
AP35	Female	Ovaries	Young	RRBL00006	35018481
AP94	Female	Ovaries	Young	RRB02962	39903075
AP34	Female	Ovaries	Young	RRB02963	27811550
AP37	Female	Ovaries	Young	RRB02960	33410348
AP71	Female	Ovaries	Young	RRBL00007	34856038
AP88	Female	Ovaries	Young	RRBL00008	38486006
AP93	Female	Ovaries	Young	RRB02961	31497198
AP55	Female	Ovaries	Old	RRB03012	37934077
AP77	Female	Ovaries	Old	RRB03013	34322656
AP80	Female	Ovaries	Old	RRB03014	36157750
AP89	Female	Ovaries	Old	RRB03015	34318423
AP141	Female	Ovaries	Old	RRB03016	33844256
AP142	Female	Ovaries	Old	RRB03017	35328097
R20	Female	Abdomen	Young	NA	NA
R29	Female	Abdomen	Young	NA	NA
R06	Male	Abdomen	Young	NA	NA
R32	Male	Abdomen	Young	NA	NA
R34	Male	Abdomen	Young	NA	NA
R07	Female	Abdomen	Young	NA	NA
R33	Male	Abdomen	Young	NA	NA
R05	Female	Abdomen	Young	NA	NA
R21	Male	Abdomen	Young	NA	NA
R28	Female	Abdomen	Young	NA	NA

Supplementary Table S1. Sample information and statistics (cont.)

Sample	Raw Base (G)	Error Rate (%)	Q20 (%)	Q30 (%)	GC content (%)
AP141	9.69	0.01	98.25	95.7	41.52
AP93	10.31	0.01	97.38	93.98	40.89
AP142	10.33	0.01	98.25	95.75	41.5
AP77	9.61	0.01	98.23	95.69	42.14
AP89	11.9	0.01	98.11	95.53	42.53
AP55	10.76	0.01	98.15	95.6	41.52
AP94	10.87	0.01	97.94	95.03	41.41
AP37	9.36	0.01	98.52	96.42	42.95
AP34	10.64	0.01	98.53	96.4	41.49
AP71	9.93	0.01	98.12	95.39	40.62
AP35	13.84	0.01	97.84	94.81	40.47
AP80	12.01	0.01	98.19	95.62	42
AP35	10.51	0.01	98.53	96.41	41.08
AP94	11.97	0.01	98.16	95.19	41.51
AP34	8.34	0.01	98.2	95.27	41.42
AP37	10.02	0.01	98.49	95.88	42.15
AP71	10.46	0.01	98.42	96.19	41.36
AP88	11.55	0.01	98.53	96.37	42.71
AP93	9.45	0.01	98.38	95.69	41.24
AP55	11.38	0.01	98.17	95.51	39.76
AP77	10.3	0.01	98.28	95.6	40.46
AP80	10.85	0.01	97.97	95.06	40.28
AP89	10.3	0.01	98.04	95.16	41.51
AP141	10.15	0.01	97.88	94.83	39.58
AP142	10.6	0.01	98	95.09	39.69
R20	NA	NA	NA	NA	NA
R29	NA	NA	NA	NA	NA
R06	NA	NA	NA	NA	NA
R32	NA	NA	NA	NA	NA
R34	NA	NA	NA	NA	NA
R07	NA	NA	NA	NA	NA
R33	NA	NA	NA	NA	NA
R05	NA	NA	NA	NA	NA
R21	NA	NA	NA	NA	NA
R28	NA	NA	NA	NA	NA

Supplementary Table S1. Sample information and statistics (cont.)

Sample	Mapped Reads(%)	Properly Paired(%)
AP141	83.40	78.16
AP93	75.27	69.38
AP142	79.84	74.32
AP77	79.71	74.03
AP89	73.41	67.05
AP55	77.95	72.08
AP94	62.79	58.07
AP37	76.98	72.13
AP34	81.84	76.85
AP71	79.86	74.15
AP35	76.68	71.58
AP80	75.17	69.39
AP35	76.76	71.96
AP94	82.03	77.30
AP34	86.45	81.59
AP37	83.17	78.42
AP71	85.70	80.81
AP88	86.99	82.85
AP93	82.80	78.53
AP55	85.29	80.51
AP77	87.83	83.17
AP80	86.56	81.74
AP89	85.34	80.68
AP141	88.09	83.44
AP142	87.19	82.62
R20	38.20	34.46
R29	31.92	29.07
R06	60.95	53.47
R32	85.00	77.32
R34	29.08	26.05
R07	83.25	76.15
R33	35.84	30.56
R05	71.44	64.81
R21	38.06	31.36
R28	61.37	54.10

Supplementary Table S1. Sample information and statistics

H. melpomene rosina mRNA sequencing and mapping statistics. Sample ID, species, tissue, stage of collection for mRNA 150bp PE directionally sequenced reads for this project. Samples mapped to *H. melpomene* genome v2.1. Walters *et al.* (2015) sample mapping statistics to *H. melpomene* genome v2.1.

Supplementary Table S2. Orthologue prediction improvement with Hmel2.1 annotation

Statistics	Values	Annotation
# genes	33137	Hmel2.0
# genes in orthogroups	24274	Hmel2.0
# unassigned genes	8863	Hmel2.0
% genes in orthogroups	73.3	Hmel2.0
% unassigned genes	26.7	Hmel2.0
# orthogroups	9320	Hmel2.0
# species-specific orthogroups	15	Hmel2.0
# genes in species-specific orthogroups	56	Hmel2.0
% genes in species-specific orthogroups	0.2	Hmel2.0
Mean orthogroup size	2.6	Hmel2.0
Median orthogroup size	2.0	Hmel2.0
G50 assigned genes	2	Hmel2.0
G50 all genes	2	Hmel2.0
O50 assigned genes	3252	Hmel2.0
O50 all genes	5468	Hmel2.0
# of orthogroups with all species present	9305	Hmel2.0
# of single-copy orthogroups	6846	Hmel2.0
# of genes	41779	Hmel2.1
# of genes in orthogroups	29698	Hmel2.1
# of unassigned genes	12081	Hmel2.1
% of genes in orthogroups	71.1	Hmel2.1
% of unassigned genes	28.9	Hmel2.1
# of orthogroups	11062	Hmel2.1
# of species-specific orthogroups	18	Hmel2.1
# of genes in species-specific orthogroups	105	Hmel2.1
% of genes in species-specific orthogroups	0.3	Hmel2.1
Mean orthogroup size	2.7	Hmel2.1
Median orthogroup size	2.0	Hmel2.1
G50 assigned genes	2	Hmel2.1
G50 all genes	2	Hmel2.1
O50 assigned genes	3638	Hmel2.1
O50 all genes	6658	Hmel2.1
# orthogroups with all species present	11044	Hmel2.1
# of single-copy orthogroups	8095	Hmel2.1

Supplementary Table S2. Orthologue prediction improvement with Hmel2.1 annotation

Statistics on orthologue prediction between *H. melpomene* v2.0 annotation and *H. erato* annotation; and on orthologue prediction between *H. melpomene* v2.1 annotation and *H. erato* annotation.

Supplementary Table S3. Mean and median read-depth for resequenced whole-genome *H. melpomene* samples

Sample	Species	Sex	Location	Mean RD	Median RD
CAM000531	*H. m. rosina*	Male	9º87'N 7º96'W	30.59	29
CAM000533	*H. m. rosina*	Male	9º87'N 7º96'W	28.83	21
CAM000546	*H. m. rosina*	Male	9º87'N 7º96'W	27.47	26
CAM001841	*H. m. rosina*	Male	9º87'N 7º96'W	28	28
CAM001880	*H. m. rosina*	Male	9º87'N 7º96'W	22.76	23
CAM002045	*H. m. rosina*	Male	9º87'N 7º96'W	25.7	26
CAM002059	*H. m. rosina*	Male	9º87'N 7º96'W	36.77	32
CAM002071	*H. m. rosina*	Male	9º87'N 7º96'W	26.43	21
CAM002519	*H. m. rosina*	Male	9º87'N 7º96'W	26.68	22
CAM002552	*H. m. rosina*	Male	9º87'N 7º96'W	26.83	22

Supplementary Table S3. Mean and median read- depth for resequenced whole-genome *H. melpomene* samples

H. melpomene resequenced samples mapped to Hmel2 genome using BWA-MEM (Davey et al., 2017).

Supplementary Table S4. dN/dS ratios from pairwise alignments for Z linked and autosomal genes

		Z			
Sp.	Sex	CDS	dN	dS	dN/dS
H.melp	All	200	0.022 [0.018; 0.028]	0.189 [0.18; 0.2]	0.120 [0.098; 0.145]
H.melp	Fem	28	NA	NA	0.120 [0.069; 0.183]
H.melp	Male	96	NA	NA	0.148 [0.122; 0.172]
H.melp	Un	193	NA	NA	0.107 [0.078; 0.143]
P.tithonus	All	90	0.025 [0.019; 0.031]	0.31 [0.26; 0.36]	0.082 [0.06; 0.10]
P.tithonus	Fem	90	NA	NA	0.10 [0.058;0.14]
P.tithonus	Male	90	NA	NA	0.066 [0.047;0.089]
P.tithonus	Un	90	NA	NA	0.079 [0.057;0.10]
M. jurtina	All	90	0.025 [0.019;0.031]	0.31 [0.26;0.36]	0.082 [0.065;0.10]
M. jurtina	Fem	90	NA	NA	0.11 [0.076;0.14]
M. jurtina	Male	90	NA	NA	0.066 [0.045;0.086]
M. jurtina	Un	90	NA	NA	0.069 [0.041;0.093]

Supplementary Table S4. dN/dS ratios from pairwise alignments for Z linked and autosomal genes (cont.)

		Autosomes			
Sp.	Sex	CDS	dN	dS	dN/dS
H.melp	All	7464	0.018 [0.017; 0.018]	0.162 [0.16; 0.17]	0.110 [0.106; 0.113]
H.melp	Fem	1231	NA	NA	0.113 [0.105; 0.121]
H.melp	Male	1238	NA	NA	0.113 [0.145; 0.167]
H.melp	Un	4739	NA	NA	0.0978 [0.093; 0.102]
P.tithonus	All	5212	0.025 [0.019; 0.031]	0.31 [0.26; 0.36]	0.094 [0.090; 0.097]
P.tithonus	Fem	922	NA	NA	0.10 [0.097; 0.11]
P.tithonus	Male	922	NA	NA	0.095 [0.089; 0.10]
P.tithonus	Un	922	NA	NA	0.083 [0.077; 0.089]
M. jurtina	All	5212	0.025 [0.019;0.031]	0.31 [0.26; 0.36]	0.094 [0.090;0.097]
M. jurtina	Fem	922	NA	NA	0.095 [0.090; 0.10]
M. jurtina	Male	922	NA	NA	0.096 [0.090; 0.10]
M. jurtina	Un	922	NA	NA	0.089 [0.083;0.096]

Supplementary Table S4. dN/dS ratios from pairwise alignments for Z linked and autosomal genes

dN/dS rations calculated from pairwise alignments for Z linked and autosomal genes. dN/dS calculated for *M. jurtina* and *P. tithonus* included for comparison to *H. melpomene*. *M. jurtina* and *P. tithonus* estimates calculated in Rousselle *et al.* (2016). Intervals represent 95% confidence intervals obtained by bootstrapping genes (1000 replicates). All – includes female, male and unbiased genes.

Supplementary Table S5. π_n/π_s ratios from pairwise alignments for Z linked and autosomal genes

		Z		Autosomes		
Species	Sex	π_s	π_n/π_s	π_s	π_n/π_s	π_{sZ}/π_{sA}
H.melpomene	All	0.012 [0.011;0.013]	0.111 [0.098;0.126]	0.027 [0.026;0.027]	0.103 [0.100;0.107]	0.444
H.melpomene	Female	0.016 [0.013;0.020]	0.0938617 [0.059;0.136]	0.025 [0.024;0.027]	0.106 [0.10;0.113]	0.64
H.melpomene	Male	0.015 [0.01;0.02]	0.136 [0.112;0.162]	0.035 [0.033;0.036]	0.127 [0.118;0.138]	0.429
H.melpomene	Unbiased	0.0106 [0.009;0.012]	0.10414 [0.09;0.125]	0.025 [0.024;0.026]	0.094 [0.091;0.098]	0.424
P.tithonus	All	0.006 [0.005;0.007]	0.131 [0.096;0.171]	0.010 [0.010;0.098]	0.126 [0.121;0.131]	0.599
P.tithonus	Female	NA	0.10 [0.049;0.16]	NA	0.137 [0.119;0.151]	NA
P.tithonus	Male	NA	0.17 [0.11;0.23]	NA	0.128 [0.120;0.137]	NA
P.tithonus	Unbiased	NA	0.12 [0.091;0.17]	NA	0.113 [0.108;0.124]	NA
M. jurtina	All	0.01 [0.009;0.012]	0.086 [0.068;0.108]	0.031 [0.030;0.031]	0.073 [0.071;0.078]	0.323
M. jurtina	Female	NA	0.044 [0.017;0.078]	NA	0.087 [0.079;0.095]	NA
M. jurtina	Male	NA	0.11 [0.082;0.15]	NA	0.082 [0.075;0.090]	NA
M. jurtina	Unbiased	NA	0.08 [0.061;0.11]	NA	0.065 [0.061;0.069]	NA

Supplementary Table S5. π_n/π_s ratios from pairwise alignments for Z linked and autosomal genes

π_n/π_s ratios calculated from pairwise alignments for Z linked and autosomal genes. π_{sZ}/π_{sA} ratio used to estimate Ne_Z/Ne_A. π_n/π_s for *M. jurtina* and *P. tithonus* included for comparison with *H. melpomene*. *M. jurtina* and *P. tithonus* ratios calculated in Rousselle *et al.* (2016). Intervals represent 95% confidence intervals obtained by bootstrapping genes (1000 replicates).

Supplementary Table S6. Rate of adaptation (α) and adaptive (ω_α) and non-adaptive (ω_{na}) substitutions rates

		Z			Autosomes		
Species	Sex	α	ω_α	ω_{na}	α	ω_α	ω_{na}
H.melp.	All	0.675 [0.647; 0.704]	0.069 [0.072; 0.066]	0.033 [0.030; 0.036]	0.629 [0.622; 0.636]	0.062 [0.061; 0.063]	0.036 [0.036; 0.037]
H.melp.	Female	0.699 [0.595; 0.811]	0.069 [0.058; 0.080]	0.029 [0.019; 0.040]	0.635 [0.620; 0.650]	0.066 [0.065; 0.068]	0.038 [0.037; 0.040]
H.melp.	Male	0.646 [0.596; 0.697]	0.090 [0.083; 0.097]	0.049 [0.042; 0.056]	0.630 [0.616; 0.646]	0.087 [0.085; 0.089]	0.051 [0.049; 0.053]
H.melp.	Unbiased	0.537 [0.500; 0.576]	0.048 [0.044; 0.051]	0.041 [0.038; 0.044]	0.538 [0.529; 0.547]	0.047 [0.046; 0.048]	0.040 [0.039; 0.041]
P.tithonus	All	0.66 [0.24;1]	0.054 [0.019; 0.094]	0.028 [0.00; 0.064]	0.63 [0.59; 0.68]	0.059 [0.053; 0.065]	0.035 [0.030; 0.038]
M.jurtina	All	0.72 [0.34;1]	0.059 [0.027; 0.094]	0.023 [0.00; 0.052]	0.78 [0.74;0.81]	0.072 [0.068; 0.077]	0.021 [0.018; 0.023]

Supplementary Table S6. Rate of adaptation (α) and adaptive (ω_α) and non-adaptive (ω_{na}) substitutions rates

α, ω_α and ω_{na} ratios obtained from pairwise alignments for Z linked and autosomal genes estimated by using the method of Eyre-Walker and Keightley (2009) as implemented in Galtier (2016). *M. jurtina* and *P. tithonus* included for comparison to *H. melpomene*. *M. jurtina* and *P. tithonus* estimates calculated in Rousselle *et al.* (2016). Intervals represent 95% confidence intervals obtained by bootstrapping genes (1000 replicates).

Supplementary Methods and Results

π_S/π_n and dNdS ratios influence on expression level by chromosome

To test whether gene expression level and chromosome number (i.e. the different chromosomes within the genome) have a significant effect on π_S/π_n and dNdS ratios I used a multiple regression analysis. I establish the linear models:

$$\log(\pi_{nij}) \sim \log(\pi_{sij}) + \text{chromosome\_number}_j + \log(\text{FPKM}_i)$$

$$\log(d_{Nij}) \sim \log(d_{sij}) + \text{chromosome\_number}_j + \log(\text{FPKM}_i)$$

using R (v3.2.5). FPKM_i is the mean FPKM of gene *i* across the 10 individuals. 477 genes with no polymorphism and 16 with no divergence were removed from the analysis. I plotted diagnostic plots of residuals versus fitted values.

Z linked genes have a median expression level significantly smaller than autosomal linked genes

Z linked genes have a median expression level significantly smaller than autosomal linked genes regardless of chromosome number (Figure SM1). Using a multiple regression approach we found that π_n and dN were significantly negatively correlated to expression level. This is true for π_n for both autosomal linked ($P < 0.01$ for all chromosomes) and Z linked genes ($P < 0.01$) and can be interpreted as increased strength of purifying selection on highly expressed genes (Figure SM2). The lack of a Z chromosome effect on π_n/π_s despite reduced expression and smaller effective population size means that there is no indication that the Z chromosome experiences reduced efficacy of purifying selection. However, this pattern was not observed for dN in the autosomes, which would in turn suggest an effect of hemizygosity on the efficacy of purifying selection ($P > 0.3$ for all chromosomes). The

diagnostic plots for the linear regression analysis of gene expression and dNdS illustrates, however, that the model is a bad fit to the data and so there are likely to be non-linear relationships in the data that are not being captured by the model (SM3).

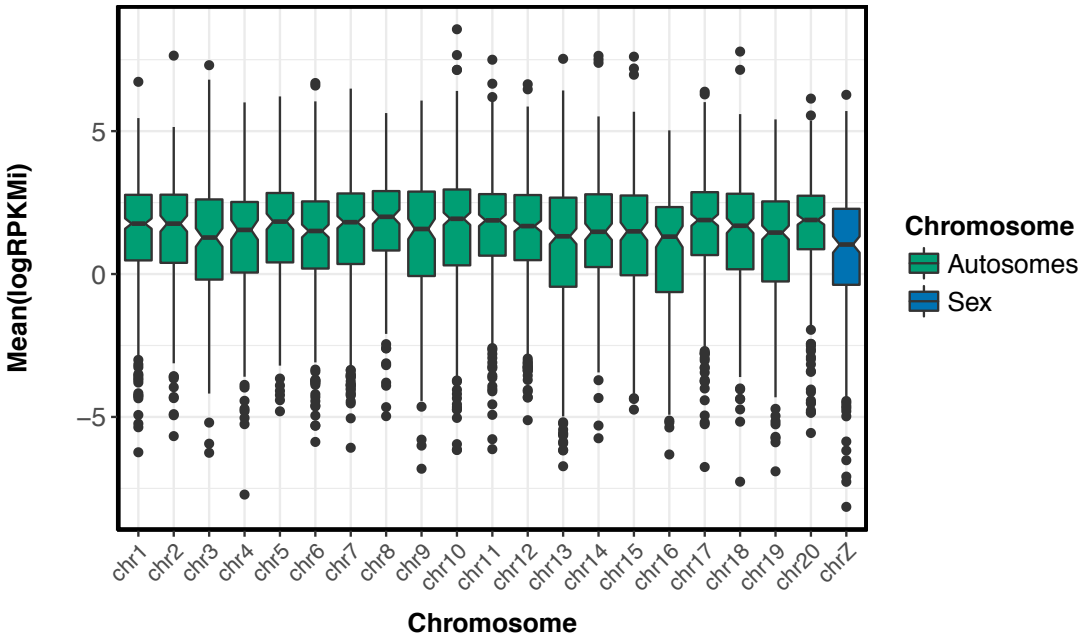


Figure SM1. Median expression level of Z and autosomal linked genes by chromosome number

Median expression level of Z linked genes is significantly smaller than autosomal linked genes ($P < 0.05$ when all the autosomes are considered separatly). Notches on boxplot display the confidence intervals around the median.

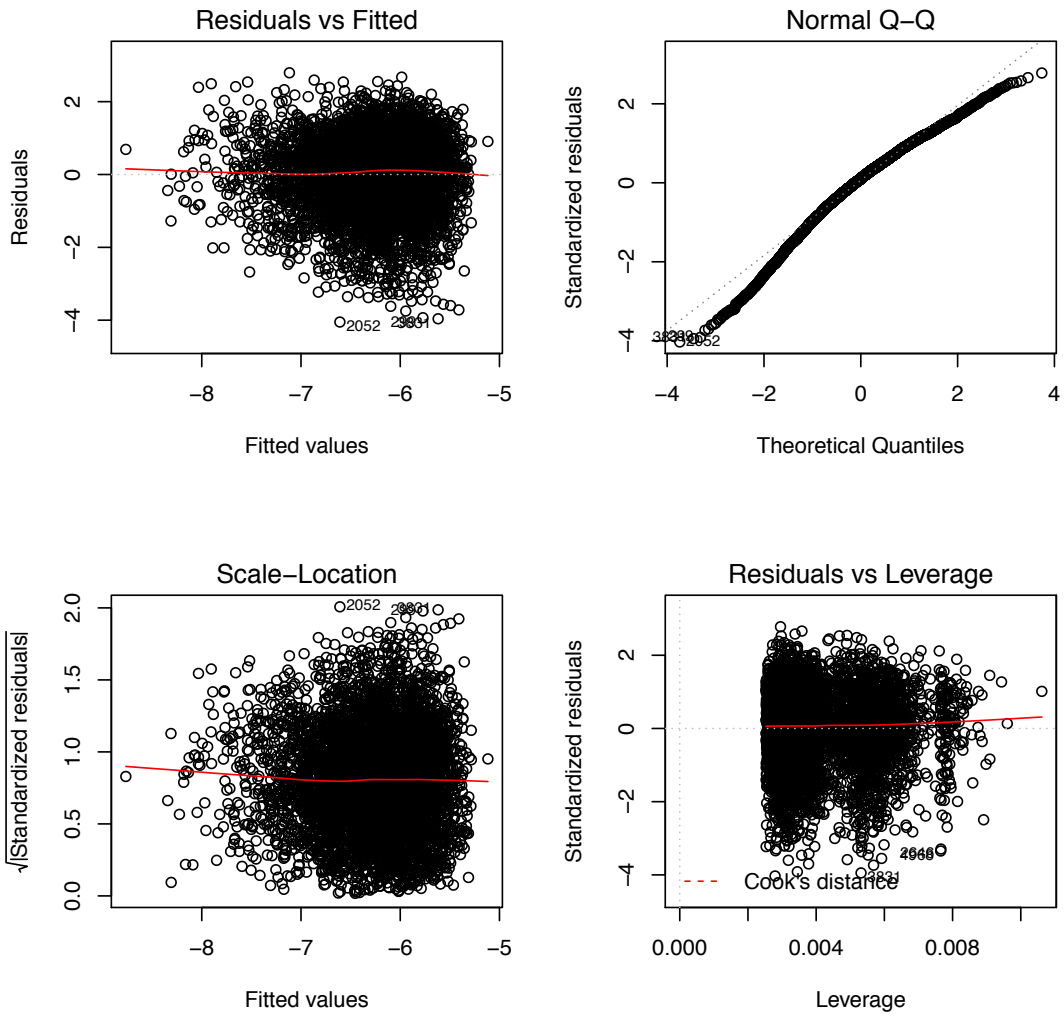


Figure SM2. π_n is negatively correlated to expression level

Multiple regression approach shows that π_n was significantly negatively correlated to expression level. Plotted *Residuals vs Fitted* shows spread residuals around the horizontal line without distinct patterns. *Normal Q-Q* follow a straight line with residuals well lined. The *Scale-Location* plot shows residuals spread equally around range of predictors. There is equal variance or homoscedasticity. *Residuals vs Leverage* plot does not identify any influential outliers in the linear regression analysis.

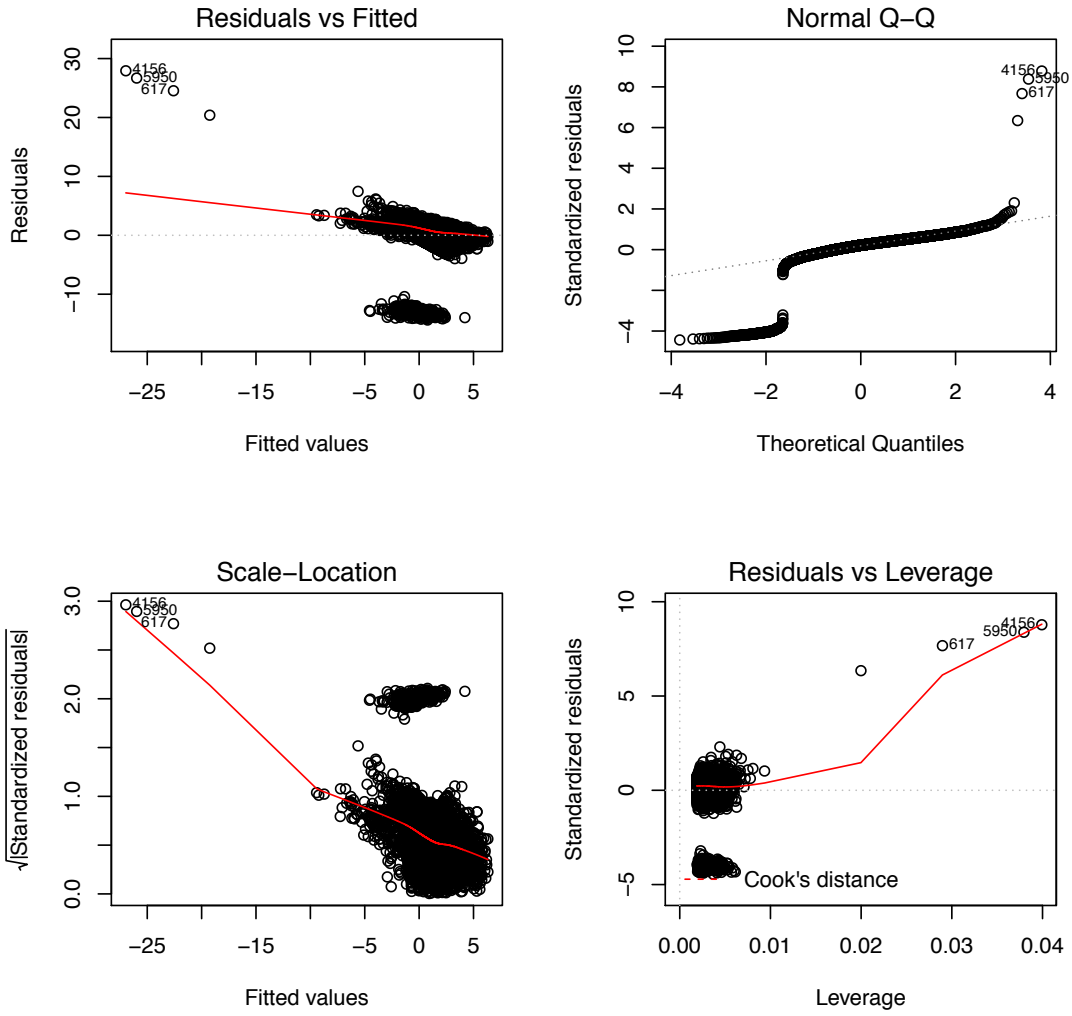


Figure SM3. dNdS and expression level the model is a bad fit to the data

There is no equal spread of the residuals around the horizontal line so there may be non-linear relationships in the data The *Residuals vs Fitted* plot shows spread residuals around the horizontal line with distinct patterns. *Normal Q-Q* do not follow a straight line. The *Scale-Location* plot shows residuals unequally spread around range of predictors. *Residuals vs Leverage* plot identifies influential outliers in the linear regression analysis that, even after being removed, do not improve the fit of the model significantly (not shown).

CHAPTER 3

Sterility in *Heliconius cydno* x *Heliconius melpomene* F1 female hybrids: a phenotypic and gene expression study of hybrid incompatibilities

Abstract

Understanding the genetics underlying speciation has been a long-standing goal of evolutionary biology. Hybridization between two species is often maladaptive and results in offspring with decreased fitness compared to the parental forms. The combination of divergent genomes within a hybrid can result in profound changes to both the genome and the transcriptome. Studying such gene expression changes can offer insights into the genetic mechanisms underlying hybrid fitness and to the evolution of reproductive isolation. Here I identify genes potentially involved in hybrid sterility by analysing ovary gene expression data from *H. melpomene*; *H. cydno*; and their F1 hybrids. These two sympatric species show low levels of inter-specific hybridisation and hybrid F1 female progeny is always sterile. Overlaps between differentially expressed genes and 1) previously identified quantitative trait loci controlling hybrid sterility; and 2) regions of reduced gene flow between the two species; are preformed to identify loci potentially involved in the species barrier.

Introduction

The process of speciation involves the evolution of barriers to gene flow between two populations. Gene flow, a homogenizing force, halts the accumulation of genetic differences between populations and so, reproductive isolation, is a vital component for the maintenance of separate species. In inter-specific crosses, post-zygotic isolation can range from hybrid breakdown, to sterility or inviability. Post-zygotic isolation can result from the interaction of at least two loci (i.e. Bateson-Dobzhansky-Muller (BDM)) (Dobzhansky, 1936; Muller 1940, 1942; Bateson, 1909). The BDM model postulates that post-zygotic isolation arises in allopatry and is a consequence of dominant negative epistasis between alleles that experienced distinctive evolutionary trajectories. The alleles that diverged separately in different populations only reduce fitness in the hybrids and are not subject to negative selection until they co-occur. Several genetic studies in animals and plants have identified and mapped the alleles underlying such hybrid phenotypes and characterized the resulting developmental defects (Presgraves *et al.*, 2003; Ting *et al.*, 2004; Brideau *et al.*, 2006; Bikard *et al.*, 2009; Phadnis, 2011; Sweigart and Flagel, 2015).

If reproductive isolation is already complete, genes that currently influence reproductive isolation may not have been involved in speciation. Studying such genes may help to understand the pathways that are responsible for phenotypic breakdown once interbreeding occurs (Nosil and Schluter, 2011). However, studies of reproductive isolation where interbreeding is ongoing, may lead to the identification of genes that are currently implicated in the maintenance of species barriers.

Recent work in speciation has focused on the importance of genomic architecture during speciation with gene flow. In order to reduce the effect of recombination breaking-down pre-mating isolation barriers and preventing further genetic divergence, different architectures can to be favoured. Theory predicts that speciation with gene flow is facilitated where: 1) genes under

divergent selection also cause reproductive isolation, for example genes causing host-preference in *Drosophila* (Matsuo *et al.*, 2007); 2) pre-mating isolation arises through the fixation of the same allele in both of the populations like it occurs in *D. persimilis*, a so-called one-allele model (Ortiz-Barrientos, 2005); or 3) there is physical linkage between the alleles experiencing divergent selection and pre-mating isolation (Merrill *et al.*, 2011; Servedio *et al.* 2003; Servedio *et al.* 2011). Identification of loci causing reproductive isolation between hybridising species can therefore help to shed light on the mechanisms underlying speciation with gene flow.

As populations diverge through the speciation process, loci under selection are more differentiated than the genome average (Nosil *et al.*, 2009). Under ongoing gene flow, “genic models of speciation”, predict restricted gene flow between a few genomic regions under divergent selection flanked by a homogeneous background of unlinked neutral loci (Emelianov *et al.*, 2004; Turner *et al.*, 2005; Lexer and Widmer, 2008; Nosil *et al.*, 2009). Both coding and non-coding variation may constitute barrier loci of elevated differentiation and contribute to pre- and post-zygotic reproductive isolation (McDermott and Noor, 2010). The democratization of next-generation sequencing technology allows to query how important these genomics regions of divergence might be during the speciation process within wild-populations (Turner *et al.*, 2005; Ellegren *et al.*, 2012; Renaut *et al.*, 2013; Martin *et al.*, 2013; Poelstra *et al.*, 2014).

Proving the existence of “genomic islands of divergence”, however, has been less straightforward than originally predicted (Wolf and Ellegren, 2017; Ravinet *et al.*, 2017). On one hand, putative genomic islands originally identified, have since been shown to actually be regions of low genic diversity that give the false appearance of high divergence and low gene flow (Cruickshank and Hahn, 2014). On the other hand, if large fractions of the genome are subject to directional selection on quantitative traits, polygenic selection can have genome-wide effects (Pritchard and Di Rienzo, 2010). Polygenic adaptation has played a pervasive role in shaping genotypic

variation in modern humans (Field *et al.*, 2016) and it pervasive polygenic selection may maintain species differences in the face of ongoing gene flow (Josephs and Wright, 2016).

Although speciation biologists typically view hybrids as unfit and forming a barrier to gene flow between species, in some cases they can also contribute useful novel variation. In hybrids, alleles from different genetic backgrounds can interact to produce novel phenotypes. More specifically, changes to gene expression in hybrids can fall outside the range of expression seen in parents (i.e. transgressive or epistatic) (Hegarty *et al.*, 2008). Studies of the molecular and genetic basis of non-additive, epistatic and transgressive interactions have contributed to our understanding of the evolution of: 1) hybrid incompatibilities (BDMI) (Dobzanky, 1937; Muller, 1940; Bateson, 1909); 2) migration load (immigration of locally maladapted alleles) (Bolnick and Nosil, 2007); and 3) adaptive phenotypes, as observed during hybrid speciation and heterosis (García-Ramos and Kirkpatrick, 1997; Jiggins *et al.*, 2008). In summary, hybridization of two different genomes within an individual provides a source of phenotypic novelty upon natural selection acts. Natural selection can facilitate the establishment of a new species when the hybrids have higher fitness than the parents (i.e. hybrid speciation, heterosis); or act to remove these individuals from the population (i.e. hybrid incompatibilities, migration load).

Studies of gene expression in hybrids can be used to understand the genetic changes that underlie hybrid phenotypes. Patterns of gene expression in hybrids can result from interactions between alleles of parental genomes and changes in regulatory and transcriptional networks (Reiland and Noor, 2002; Wittkopp *et al.*, 2004; Ranz and Machado, 2006; Wittkopp *et al.*, 2008; McManus *et al.*, 2010). This variation in mRNA abundance is the result of: 1) changes that affect genes in *cis* (including associated regulatory sequences); 2) changes in the activity of *trans*-acting factors; or 3) both (Landry *et al.*, 2007).

Although typically gene expression differences between species might result from divergent natural selection, it is also the case that lineages can accumulate genetic differences in orthologous genes and still exhibit similar phenotypes. Molecular coevolution of gene sequences guarantees that gene function is maintained despite the accumulation of nucleotide differences in their regulatory and coding sequences (Landry *et al.*, 2007). Studying gene expression differences in hybrids compared to parent species can therefore provide insights into the genetic mechanism by which hybridization leads to non-additive, transgressive or epistatic segregation and its link to hybrid fitness. Ultimately, the identification of such genes is a step towards understanding the evolution of reproductive isolation.

Heliconius are neotropical butterflies best known for their Müllerian mimicry. Studies of *Heliconius* have contributed to answering evolutionary questions covering a broad range of research topics; from phylogenetics to ecology, behaviour, and genetics (Merrill *et al.*, 2015). In particular, Heliconius have been studied to understand the process of speciation. Heliconius species are typically separated by a wide range of barriers to gene flow including pre-zygotic (host-plant and mate-preference) and post-zygotic (sterility and increased predation) barriers (Jiggins *et al.*, 2001; Naisbit *et al.*, 2002; Merrill *et al.*, 2012; 2013).

Heliconius butterfly hybrids often show pronounced sex-specific asymmetries in fitness, following Haldane's rule, which states that when one sex in the progeny of a cross between two groups is inviable or sterile, it is typically the heterogametic sex that is affected (Haldane, 1922). Supporting Haldane's predictions the female *H. cydno* x *H. melpomene* hybrids are sterile but the male progeny of the same cross shows no signs of decreased fitness (Jiggins *et al.*, 1996; Naisbit *et al.*, 2001). In fertile females, the basis of oogenesis is comparable between *Heliconius* and other holometabolus insects with similar life spans. For holometabolus insects, oogenesis can be divided into three stages: 1) pre-vitellogenesis, where the nurse cells develop; 2) vitellogenesis, where the oocyte grows rapidly through the accumulation of yolk; and 3)

choriogenesis, when the eggshell forms after the secretions of hormones such as the juvenile hormone. In each ovariole there is an array of developing follicle cells: starting with the dividing germ cells in the germanium, and finishing with the mature oocyte ready for fertilization in the common oviduct (Dunlap-Pianka *et al.*, 1977) (Figure 1).

Asymmetric post-mating isolation occurs when the strength of reproductive isolation between taxa differs between reciprocal crosses. Asymmetric post-mating isolation was first formalised by Darwin (1859) and it is a common phenomenon (Turelli and Moyle, 2006). The asymmetry is usually the result of BDMIs involving uni-parentally inherited genetic factors that act simultaneously with bi-directional BDMIs between autosomal loci affecting reciprocal crosses equally. By modelling both classes of two-locus BDMIs Turelli and Moyle (2006) found systematic interspecific differences in relative rates of evolution for autosomal and non-autosomal loci and concluded that unidirectional BDMIs involving sex chromosomes, cytoplasmic elements or maternal effects were likely to have an important role for the evolution of post-mating isolation barriers (Turelli and Moyle, 2006).

Here I study the process of oogenesis and its breakdown in sterile Heliconius hybrids. With well documented species incompatibilities and many species pairs at different levels of divergence *Heliconius* is an excellent system to investigate the possible genetic causes of hybrid female sterility (Jiggins *et al.*, 2001; Martin *et al.*, 2013; Kozak *et al.*, 2015). Here I focus on *H. cydno* and *H. melpomene*, two hybridising sympatric species that differ in their ecology, mimicry patterns and mate preferences. They show low levels of inter-specific hybridisation that nonetheless results in genome-wide signatures of admixture. Hybrid F1 female progeny of the *H. cydno* x *H. melpomene* cross are always sterile but an unresolved question remains over the number and identity of the genomic regions that contribute to their speciation.

In an attempt to elucidate post-zygotic isolation barriers between the two species, I analyse gene expression profiles and describe expression differences between the fertile *H. cydno* and *H. melpomene* females and the F1 sterile females. I start by examining whether these F1 females develop ovarian tissue by characterising ovarian structures of hybrids as compared to *H. melpomene* and *H. cydno* females. Then I quantify gene expression through the analysis of mRNA sequencing data from ovaries of fertile (*H. cydno* and *H. melpomene*) and sterile (hybrid) samples. Finally, I identify genes showing differential expression in hybrids that are potentially involved in hybrid sterility.

Material and Methods

Intra- and inter-specific crosses of *H. cydno* and *H. melpomene*

For this study I focus on hybrid female sterility of a hybrid cross between a *H. cydno* female and a *H. melpomene* male. While it would be valuable to perform the experiment using the reciprocal cross (*H. melpomene* female x *H. cydno* male) to address asymmetry postzygotic isolation, this was not possible. In a fashion similar to what it observed between *D. melanogaster* and *D. simulans* (Carracedo *et al.* 1998), between *H. cydno* and *H. melpomene*, we have only been able to perform hybrid crosses in one direction. Crosses were carried between *H. cydno chioneus* and *H. melpomene rosina* at the Smithsonian Tropical Research Institute insectaries in Gamboa, Panama (9º08'N 7º42'W). All mothers of broods are virgin insectary-bred females. Father of broods are wild caught individuals collected along Pipeline Road in the Soberanía National Park (9º87'N 7º96'W).

Intra-specific crosses were carried out between virgin *H. cydno* insectary bred females and wild males; and between virgin *H. melpomene* insectary bread females and wild males. Virgin *H. cydno* and *H. melpomene* females were placed at time of emergence in a mature wild male cage (4x4x2m) of the

same species. I monitored the male cage every hour. If a mating was recorded, the females were placed to lay eggs in a large cage (4x4x4m) with 5 to 8 other mated females of the same species. If a mating was not recorded within the first 48h the females were preserved.

Inter-specific crosses were carried out between virgin insectary bred *H. cydno* females and wild *H. melpomene* males. For inter-specific crosses I placed pupae from the *H. cydno* stock in a cage (4x4x4m) with 3 to 5 wild *H. melpomene* males. After eclosion, if the butterfly was female, I checked the cage every hour for matings. I left newly eclosed *H. cydno* females for 8h in the wild *H. melpomene* male cage or until a mating occurred. When an inter-specific mating occurred, the *H. cydno* females were placed to lay eggs alone in a cage (4x4x2m). When a mating did not occur within the first 8h the female was moved to a *H. cydno* male stock cage to attempt an intra-specific mating.

I collected eggs every day for both inter-specific and intra-specific crosses and laying females had access to *Psiguria* flowers; *Lantana camara*; and artificial feeders containing a 20% sugar-water solution with 5% added commercial pollen, changed every other day. All females were also provided with *Passiflora* plants with fresh shoots for laying: *H. melpomene* - *P. menispermifolia*; *H. cydno* - *P. edulis*, *P. vitifolia* and *P. williamsi*. I kept the collected eggs separated into individual plastic pots to prevent cannibalism. After hatching, I moved the 1st instar caterpillars to rearing cages (2x2x1m) with *Passiflora* plants: *H. melpomene* - *P. menispermifolia*; *H. cydno* - *P. edulis*, *P. vitifolia* and *P. williamsi*. Larvae were reared in cages until 5th instar. At 5th instar I removed the caterpillars and took them to the laboratory in large individual containers where they were allowed to pupate and emerge at a constant temperature (24-25ºC). The pupating containers in the laboratory were monitored several times a day. When a female emerged I either: 1) took it back to the insectaries and allowed to mature (Phenotypic study); or dissected it for gene expression analysis of ovary tissue (Supplementary Table S1).

Phenotype scoring of mature fertile *H. cydno* and *H. melpomene*; and *H. cydno* x *H. melpomene* sterile females

Females were mated with males of the same species following the protocol above for intra-species matings. The females were than individually marked and kept with other females of the same species for 20 days under the conditions previously described.

After 20 days each female was dissected and the phenotype of its reproductive tract was scored (Table 1). 26 out of 28 dissections were carried out in RNAlater. For those 26 dissections each female butterfly has a sample of the gut, the abdomen shell, the thorax, the bursa, and the ovaries stored in separately in RNAlater at -80ºC (Appendix B, Protocol for dissections of the reproductive tract for total RNA extraction, Figure 1). For the 2 dissections carried out in PBS the tissues were discarded after scoring the phenotype. All samples used for phenotype data collection have pictures of their reproductive tracts. The ovaries are scored both quantitatively (Swevers & Iatrou 2003) and qualitatively (Luiz 2008). For each mature female I have recorded: 1) the number of ovarioles; 2) the different stages of development of the oocytes, by category; 3) the number of vitellogenic oocytes; 4) the number of choriogenic oocytes; 5) the phenotype of the bursa copulatrix (Figure 1, Figure 2).

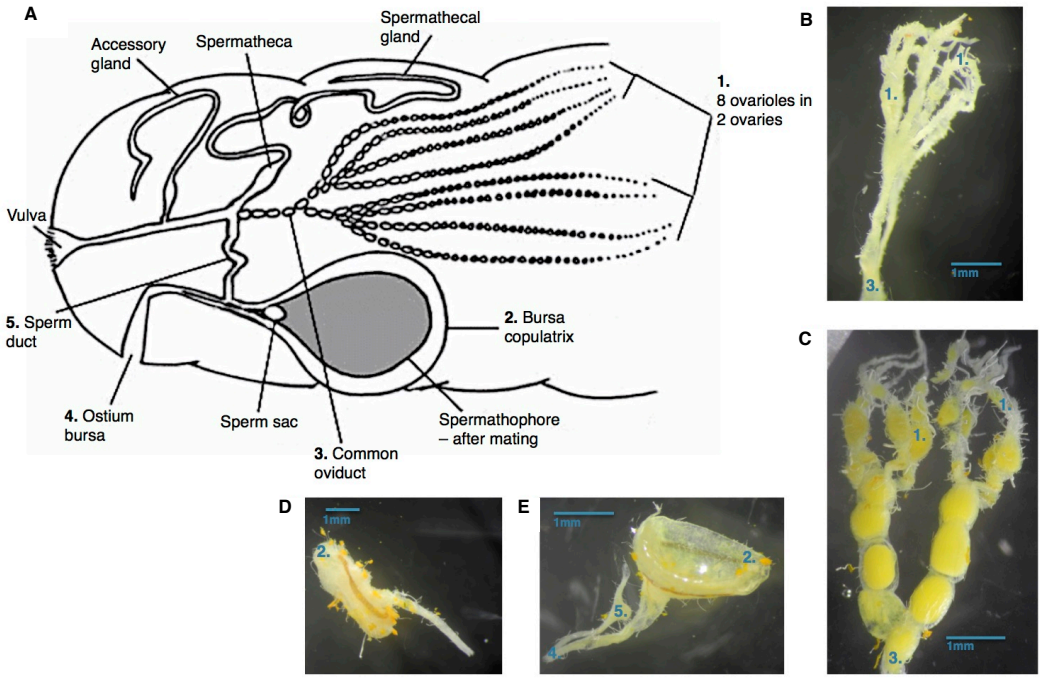


Figure 1. The reproductive tract of *Heliconius* butterflies

A. Diagram of the reproductive tract (adapted from Monarch Lab, www.monarchlab.org). **B.** Ovaries of a newly emerged virgin female with the ovaries and the ovarioles (1.) and the common oviduct (3.) labeled. **C.** Ovaries of a 20 day old mated fertile female. Ovaries and the ovarioles (1.) and the common oviduct (3.) labeled. **D.** Bursa copulatrix of a newly emerged virgin female. **E.** Bursa copulatrix of a mated mature female. Note how the digestion of the spermathophore made the bursa concave. Ostium bursa (4.) and sperm duct (5.) labeled.

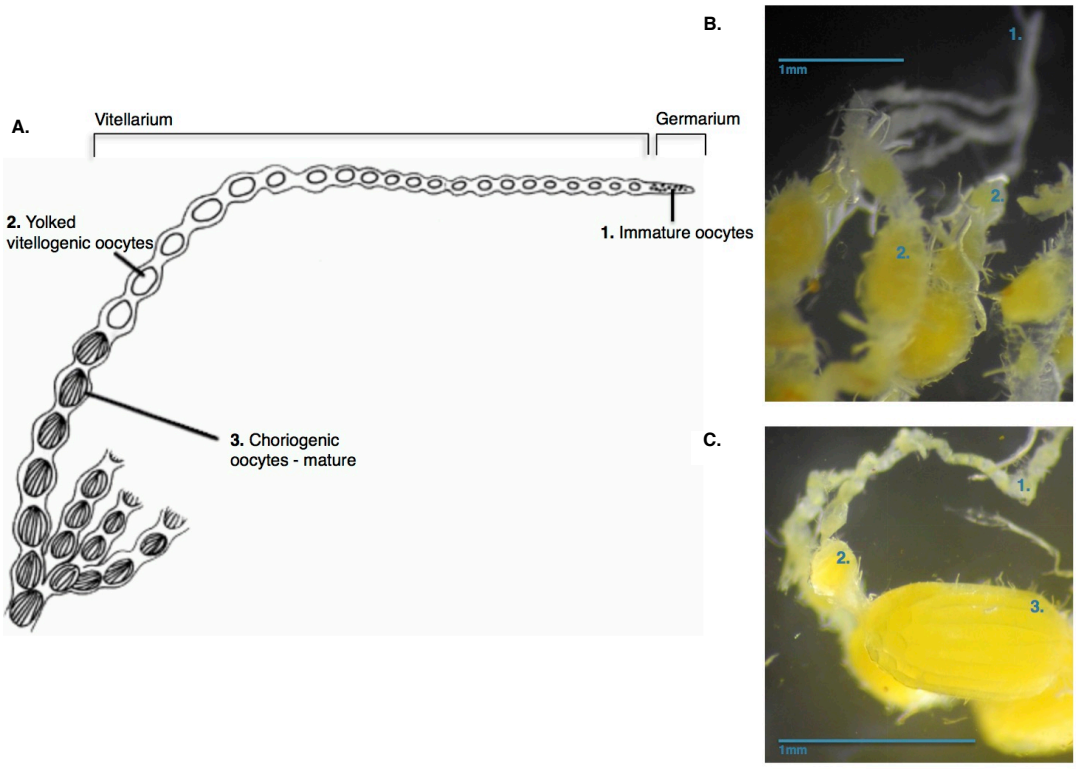


Figure 2. Qualitative and quantitative phenotype scoring of mature female ovaries

A. Diagram of oogenesis (adapted from Monarch Lab, www.monarchlab.org). **B.** Close up picture of the germanium and the vitellarium on 1 ovary (4 ovarioles) of a fertile mature *Heliconius* female. **C.** Close up picture of the germanium, vitellarium and a choriogenic oocyte on 1 ovariole of a fertile mature *Heliconius* female. In **1.** it is possible to start distinguishing the different developing oocytes. Oocytes with this phenotype were not accessed quantitatively. In **2.** the developing oocyte has undergone vitellogenesis – yolk deposited on the oocyte. In **3.** it is possible to distinguish the shell (vertical ridges) which indicates the oocyte has completed the last stage of oogenesis, choriogenesis.

Samples		Ovary 20 days after eclosion
	Phenotype scoring	4 x
	Phenotype scoring	4 x
	Phenotype scoring	8 x

Table 1. Samples used in the phenotype study

Total number of samples used to score ovary phenotype in *H. cydno*, *H. melpomene* and *H. cydno* X *H. melpomene*. All samples used to score the ovary phenotypes were dissected 20 days after eclosion.

H. cydno, H. melpomene and F1 female hybrid tissue collection to quantify gene transcript abundance

I dissected ovary tissue 1-3h after eclosion for *H. cydno, H. melpomene* and F1 hybrid females that pupated and emerged in the laboratory. The ovary tissue from young females was sequenced to quantify gene transcript abundance. I also dissected 20-days old *H. melpomene* ovary tissue to quantify gene transcript abundance. I took pictures of the reproductive tract in over 90% of the dissections performed. I carried out dissections at 24-25ºC in RNAlater (ThermoFisher, Waltham, MA) under a dissection microscope; and stored the tissue also in RNAlater at -20ºC (ThermoFisher, Waltham, MA). For each female I have a sample of the gut, the abdomen shell, the thorax, the bursa, and the ovaries stored in separate tubes in RNAlater at -80ºC (Appendix B, Protocol for dissections of the reproductive tract for total RNA extraction, Figure 1, Table 2, Figure 3).

Samples	Ovary	
	3h after eclosion	20 days after eclosion
 ~34M reads/sample	7 x 150 PE	6 x 150 PE
 ~34M reads/sample	7 x 150 PE	NA
 ~34M reads/sample	10 x 150 PE	NA

Table 2. Samples used in the gene expression study

Total number of samples used to estimate gene expression differences between *H. cydno*, *H. melpomene* and *H. cydno* X *H. melpomene*. Average read number for each group is reported as well as time of dissection: 24 samples were dissection ~3h after eclosion and 6 samples 20 days after eclosion. PE refers to paired-end RNAseq reads.

Total RNA extraction for mRNA sequencing

For 7 young *H. melpomene* ovaries, 7 young *H. cydno* ovaries, 10 young hybrid ovaries and 6 mature *H. melpomene* ovaries total RNA was extracted with a combined guanidium thiocyanate-phenol-chloroform and silica matrix protocol using TRIzol (Invitrogen, Carlsbad, CA), RNeasy columns (Qiagen, Valencia, CA) and DNaseI (Ambion, Naugatuck, CT) (Appendix C, Total RNA extraction protocol for mRNA sequencing). Total RNA integrity was checked using the Bioanalyzer RNA Nano kit (Agilent, Santa Clara, CA, USA) and NanoDrop Nucleic Acid Quantification (ThermoFisher, Waltham, MA, USA). mRNA isolated from total RNA via poly-A pull-down, directional cDNA libraries and 150bp paired-end sequencing done in Novogene Bioinfomatics

Technologies (~30M reads/sample) (Hong Kong, China) (Supplementary Table S1).

Heliconius cydno guided assembly and annotation transfer

In order to have a reference assembly for both of the species of interest, I first generated a reference-guided assembly for *H. cydno*. A *H. cydno* male Illumina assembly was available based on a method known as trio-sga (Malinsky et al. 2016). trio-sga allows the generation of a haplotypic assembly from deep-sequenced trios: a *H. cydno* mother, father and progeny (Davey *et al.,* 2017). I used progressiveCactus to align the *H. cydno* haplotypic Illumina assembly to the chromosomal version of the *H. melpomene* genome (Paten, Diekhans, *et al.*, 2011; Paten, Earl, *et al.*, 2011; Davey *et al.*, 2016) 5. The HAL database created by progressiveCactus was loaded to Ragout to produce the final reference-guided assembly (*H. cydno* reference fasta file; ordering information and unplaced scaffolds available from https://www.dropbox.com/sh/5krc7kn3u0oviwj/AADHTlQsoxQCnqZnivatNdRba?dl=0). The final *H. cydno* guided assembly has 58 scaffolds; with 22 191 fragments used for the assembly. The sum of length of the fragments used in the assembly is 261 056 210 bp. There are 13 774 unplaced fragments (17 915 151 bp, 6.63% total) and there are 7843 935 (3%) Ns introduced. The assembly N50 is 13 724 118.

I transferred the updated *H. melpomene* annotation to the *H. cydno* assembly (methods to improve the completeness of the *H. melpomene* annotation described in Chapter 2, “Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*”). I used EMBOSS Seqret (v6.6.0.0) to convert the *H. melpomene* annotation file to the embl format (Rice *et al.*, 2000). Then I used RATT to transfer the *H. melpomene* annotation (reference) to the guided *H. cydno* genome (query). RATT is part of PAGIT, a post-assembly genome-improvement toolkit (v1.0) (Swain *et al.*, 2012). I searched for synteny between the reference and the

query using MUMmer (v4.0) and detected possible errors such as start and stop codons or frameshift mutations (Kurtz *et al.*, 2004). After correcting such errors with the RATT pipeline the annotation transfer to *H. cydno* was complete (Otto *et al.*, 2011).

Read mapping, counting and estimation of variance-mean dependence

I trimmed mRNA-seq reads with default settings using Trim Galore to remove 1) adapter sequences; and 2) low quality reads from the RNA-seq mate pairs (N>10%; Qphred<5 in over 50% reads) (https://github.com/FelixKrueger/TrimGalore). Adapter sequences: 5' adapter: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT; 3' adapter: 5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG.

In order to assess mapping quality against each genome reference, I used HISAT2 (Kim *et al.*, 2015) to align read-pairs from each sample to: 1) the *H. melpomene* reference genome, 2) the *H. cydno* reference genome, 3) the trio-sga haplotypic assembly for *H. melpomene*, and 4) the trio-sga haplotypic for *H. cydno* assembly using different mapping parameters. Ultimately, I wanted to maximize the number of genic features counted for the analysis. For each sample, using each one of the 4 different genome annotations I changed the L0_x command line option to vary the maximum number of ambiguous characters allowed in the read as a function of the read length; and the rgx_x command line option to modify the length of the read gap allowed and corresponding penalties. Summary mapping statistics were calculated using samtools flagstat (v1.2) (Li *et al.*, 2009) (Figure 6). htseq-count was used to count how many aligned sequencing reads mapped to each genic feature (HTSeq v0.6.1; python v2.7.10; option: -m union) for the read-pairs mapping

to the *H. cydno* annotation and to the *H. melpomene* annotation (Anders *et al.*, 2015) (Figure 7).

Estimation of variance-mean dependence from the count data was performed with the DESeq2 (v1.14.1) of Bioconductor (v3.4) in the R software environment (v3.2.5) using the constructor function DESeqDataSetFromHTSeqCount(design=~batch+species) for all the samples using both: 1) the *H. cydno* reference genome and reference annotation; and 2) the *H. melpomene* reference genome and annotation. All the result tables were built using the DESeq2 results() function (options: betaPrior=false, test=Wald) (Love *et al.*, 2014). I filtered the results as in Walters *et al.* (2015) with log2 fold significance threshold l> 1.5l and FDR < 0.05 (options: lcfThreshold=1.5, altHypothesis="greaterAbs", alpha=0.05) (Walters *et al.*, 2015) (Figure 11). Differential expression was estimated as a combined analysis on all treatment groups and contrasts were extracted for each one of the groups being analysed.

Predicting biological processes, cellular components, molecular function and protein class for the differentially expressed genes

I used InterProScan (v5.18.57.0) (options –t n –goterms) to scan genic sequences from the differentially expressed genes against the InterPro signatures. InterPro signatures are predictive models provided by several different databases such as Gene3D, InterPro, Pfam, PRINTS, SUPERFAM, PROSITE and PANTHER. This allowed for functional analysis of proteins by classifying them into families and predicting domains and important sites (Mitchell *et al.*, 2015). I analysed the PANTHER database IDs, which can be used to infer the function of some uncharacterized genes based on their evolutionary relationships to genes with known functions (Mi *et al.*, 2016). I ran the PANTHER enrichment test on the set of differential expressed gene and their respective predicted biological functions using the *D. melanogaster* genome as the reference list with Bonferroni correction for multiple testing.

Narrowing down candidate list of gene putatively involved in reproductive isolation between *H. cydno* and *H. melpomene*

To investigate the list of candidate genes putatively responsible for the sterility phenotype of the F1 females I followed several different strategies and focused in detail on those genes that: 1) have predicted development and reproduction functions; 2) map to a previously identified sterility QTL between backcrossed *H. cydno* x *H. melpomene*; 3) map to regions of the genome where there is no gene flow between *H. cydno* and *H. melpomene*. For all the genes that overlapped any of these categories I also check whether or not they map to the duplications I identified in Chapter 1, "The comparative landscape of duplication in *Heliconius melpomene* and *Heliconius cydno*" and also, when possible, what their rate of adaptive evolution is (Chapter 2, "Lack for the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*" for details).

Constructing updated reference genome for *H. cydno* and *H. melpomene cis-* and *trans*-expression difference analysis

The *H. cydno* mothers and the *H. melpomene* fathers of the two hybrid broods (brood N1 and N9) were preserved in 2ml of 20% DMSO and 0.25M EDTA (pH 8.0) and stored at -20ºC. For the four adults (2x *H. cydno* mothers and 2x *H. melpomene* fathers), the thorax was dissected away from the head and abdomen and cut in half along the median plane. One half of thorax was used for DNA extraction with the remaining tissue returned to storage. All dissections were performed with a new sterile scalpel, fresh Parafilm, and tweezers washed in 80% ethanol. DNA for the 4 samples was extracted with the QIAGEN DNeasy Blood & Tissue kit (69504) following manufacturer's instructions for animal tissue. WGS library preparation was done in Novogene Bioinfomatics Technologies (150bp paired-end reads) and it was also

sequenced at Novogene Bioinfomatics Technologies using the Illumina HiSeq2500 (~70M reads/sample) (Hong Kong, China) (Table 3).

Sample	Species	Total # reads	Mean RD	Brood Parent
CAM25104	H. melp.	6242206307	22.76	N9
CAM25137	H. cyd.	6863171374	25.03	N9
CAM25004	H. melp.	7048554112	25.7	N1
CAM25091	H. cyd.	7051197429	25.71	N1

Table 3. WGS summary statistics for parents of broods

Summary statistics of the 4 brood parents sequenced in Illumina 2500 150 paired-end reads. Mean RD is the mean read-depth of the re-sequenced sample

The *H. cydno* mothers fastq reads were aligned to the *H. cydno* Ragout transfer reference genome and *H. melpomene* samples to the *H. melpomene* reference genome (v2.0) (Davey *et al.*, 2016) with Stampy (v1.0.23) (Lunter and Goodson, 2011) using default values for all parameters except the substitution rate, which was set to 0.01. Picard (v1.128) (picard.sorceforge.net) was used to convert SAM/BAM files and remove PCR duplicate read pairs. SNPs were called for each individual using the GATK HaplotypeCaller and combined into one final VCF file using GATK GenotypeGVCFs with options --annotateNDA and --max_alternate_alleles 30. Statistics on VCF files were calculated using VCFtools v0.1.11 (Danecek *et al.* 2011). Bcftools (v1.3) (Li *et al.*, 2009) and bedtools (v2.20.1-13- g9249816) (Quinlan and Hall, 2010) were used to process BAM and VCF files. The SNPs from each sample were used to generate 4 different alternative reference

fasta files using GATK FastaAlternateReferenceMaker: 1) *H. melpomene* alternative reference fasta brood N1; 2) *H. cydno* alternative reference fasta brood N1; 3) *H. melpomene* alternative reference fasta brood N9; and 4) *H. cydno* alternative reference fasta brood N9. The main source of error when measuring allele specific expression results from poor mapping of the RNAseq samples to the reference. My crosses are not inbred and neither were the reference genome strains and so it was necessary to have reference genomes with the same SNPs as the hybrid *H. cydno* X *H. melpomene* females to reduce erroneous counting.

Assigning *H. melpomene* X *H. cydno* reads to genes and species for *cis*- and *trans*-expression difference analysis

HISAT2 (Kim *et al.*, 2015) was used to align fastq reads from each *H. cydno* X *H. melpomene* female hybrid to both their parents' alternate fasta reference file. Each parent's alternate fasta reference files has the specific parental haplotype. I filter the hybrid mapping files so that there were no read-mismatches and all the reads were properly paired. Then, for each sample, reads that mapped equally well (no mismatches, i.e. no informative species specific SNP) to both the *H. melpomene* and the *H. cydno* parent were discarded. After filtering using these parameters there were 2 different files for each hybrid female sample: 1) reads that belong to the *H. cydno* mother allele; 2) reads that belong to the *H. melpomene* father. htseq-count was used to count how many aligned sequencing reads for each sample, and each parental species, mapped to each genic feature (HTSeq v0.6.1; python v2.7.10; option: -m union) (Anders *et al.*, 2015). I mapped species-specific reads for each sample to either the *H. cydno* reference annotation or to the *H. melpomene* reference annotation.

I tested for evidence of *cis*- and *trans*- divergence in the hybrid dataset. First, the hybrid data was analysed for evidence of differential expression as described in the *Material and Methods* section: Read mapping, counting and

estimation of variance-mean dependence. Any significant difference in the abundance of *H. cydno* and *H. melpomene* alleles between *H. cydno* and *H. melpomene* was considered evidence of expression divergence, and any significant difference in abundance of *H. cydno* and *H. melpomene* alleles in the female hybrids was considered evidence of *cis*-regulatory divergence. Genes that were differentially expressed in either: 1) the *H. cydno* and *H. melpomene* samples, or 2) in the hybrid samples were analised for *trans*-regulatory differences by comparing species specific read abundance ratios between the *H. cydno* and *H. melpomene* samples and the hybrid samples (Fisher's exact test). I calculated *cis*-regulatory divergence as the $\log_2$ transformed ratio of reads mapping to *H. melpomene* and *H. cydno* in the hybrid sample; and *trans*-regulatory divergence as the difference between $\log_2$ transformed ratios of species specific reads in the parental (i.e. *H. cydno* and *H. melpomene*) and hybrid (*H. cydno* X *H. melpomene*) samples.

Results

F1 hybrid females have less oocytes but still develop ovary structures

To the able to fully investigate the possible genetic causes of reproductive isolation in *H. cydno* x *H. melpomene* female hybrids it was necessary to measure and score the ovary phenotypes of mature F1 hybrid females, *H. cydno* and *H.* melpomene. Throughout the text the following notation will be used to refer to *H. cydno*, *H. melpomene* and *H. cydno* x *H. melpomene* samples. *H. cydno*: CPCP; *H. melpomene*: MPMP and *H. cydno* X *H. melpomene*: CPMP.

Of the total 28 females I dissected at 20-days, I used 16 to score the ovary phenotypes (Table 4, Figure 2, Figure 3). The other 12 females were not used

to score the phenotypes due to poor dissection quality. All the dissected samples were photographed.

Genotype	Sample	Nr of ovarioles	Ovary phenotype	Nr of vitellogenic eggs	Nr of choriogenic eggs	Bursa
CPCP	AP73	8	ABCD	15	13	Mated
CPCP	AP74	6	ABCD	9	7	Mated
CPCP	AP76	8	ABCD	16	11	Mated
CPCP	AP79	8	ABCD	17	7	Mated
MPMP	AP77	8	ABCD	16	5	Mated
MPMP	AP80	8	ABCD	14	13	Mated
MPMP	AP89	8	ABCD	25	11	Mated
MPMP	AP141	8	ABCD	14	8	Mated
CPMP	AP56	8	ABCD	7	4	Mated
CPMP	AP75	8	ABCD	9	12	Mated
CPMP	AP92	8	E	0	0	Mated
CPMP	AP103	8	BD	2	1	Mated
CPMP	AP104	8	E	0	0	Mated
CPMP	AP105	8	B	2	0	Mated
CPMP	AP106	8	E	0	0	Mated
CPMP	AP107	8	C	1	0	Mated

Table 4. Phenotypic scoring of mature fertile (*H. cydno* and *H. melpomene*) and sterile (*H. cydno* x *H. melpomene*) females

Counts of number of ovarioles, number of vitellogenic eggs and number of choriogenic eggs in 20 day old *Heliconius* females. Ovary phenotype classification adapted from Luiz et al. (2008). A – Oocytes are at the pre-vitellogenic stage; B – Ovarioles have early vitellogenic oocytes; C – Ovarioles with late vitellogenic oocytes; D – Ovarioles with choriogenic oocytes; E – No oocytes (Luiz, 2008).

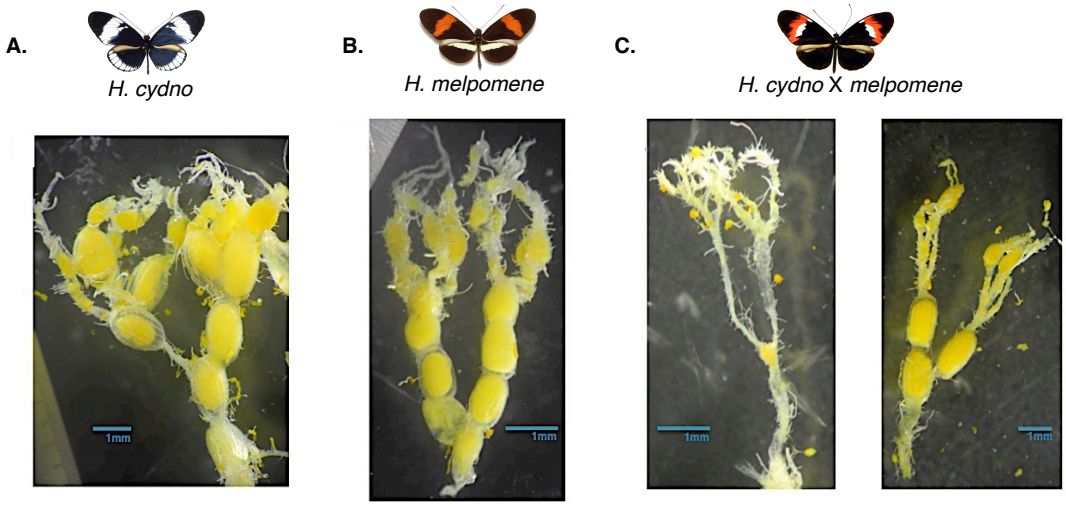


Figure 3. Example of different ovary phenotypes observed for mature fertile (*H. cydno* and *H. melpomene*) and sterile (*H. cydno* x *H. melpomene*) females

A. and **B.** have oocytes at all stages of development (ovary phenotype = ABCD). **C.** Illustrates the different phenotypes observed in the unfertile but mature and mated female offspring of two intra-specific (CP x MP) crosses. Both crosses had some CPMP females that were able to produce oocytes (ovary phenotype = BD) and some not (ovary phenotype = E). No CPMP mated female from this analysis laid eggs.

There are significant differences between the number of oocytes at each stage between the fertile *H. cydno* and *H. melpomene* and the F1 hybrids.

Mean oocyte counts are significantly different between the sample groups (Figure 4) (Swevers and Iatrou, 2003).

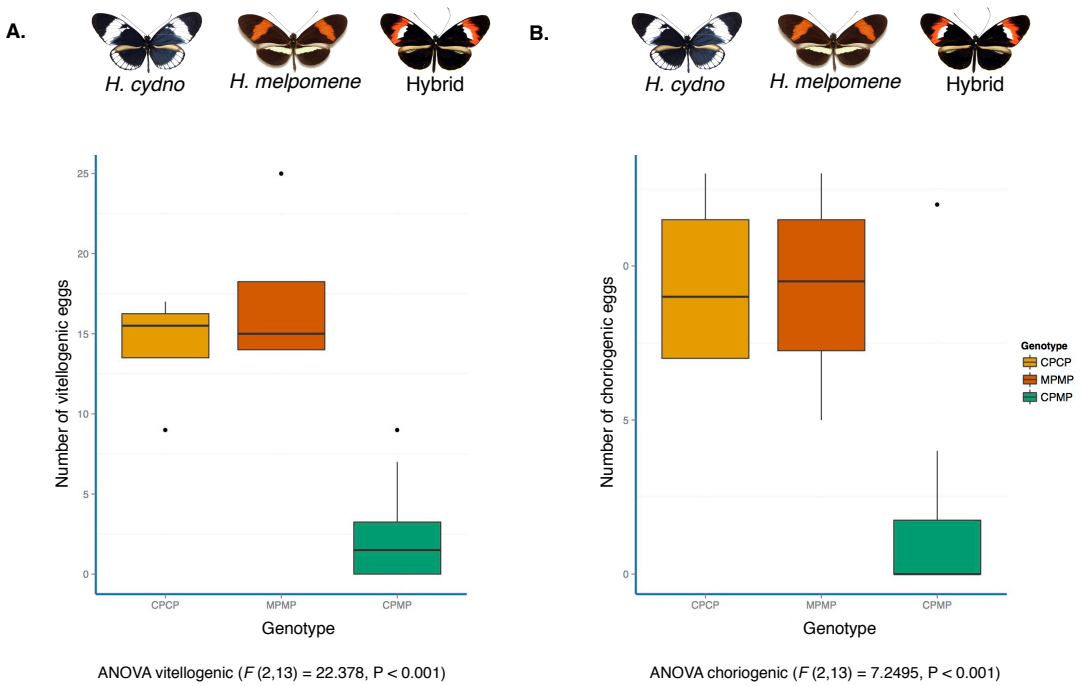


Figure 4. Distributions of vitellogenic and choriogenic oocytes counted for the three sample groups

A. Number of vitellogenic oocytes and **B.** number of choriogenic oocytes counted in *H. cydno*, *H. melpomene* and F1 hybrids sample groups. I determined statistical significance between groups by one-way ANOVA for the vitellogenic ($F(2,13) = 22.378$, $P < 0.001$) and for the choriogenic ($F(2,13) = 7.2495$, $P < 0.001$) stages.

I conducted a post-hoc Tukey's HSD test and confirmed that the significant differences arose from the comparison between the fertile (*H. cydno* and/or *H. melpomene*) and the sterile (*H. cydno* x *H. melpomene*) females, and that the

mean differences between *H. cydno* and *H. melpomene* females are not statistically significant (Crawley 2005) (Figure 5).

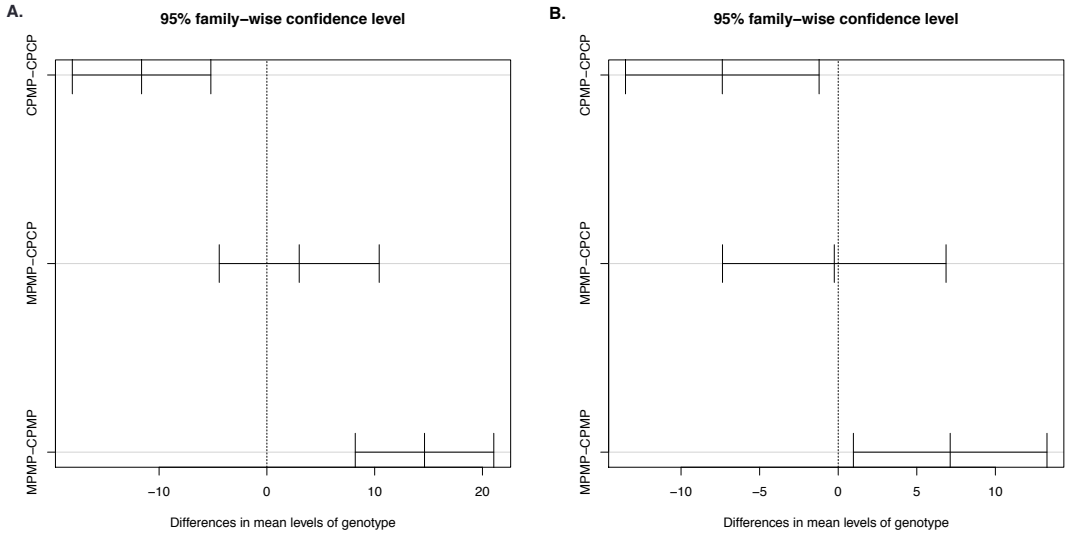


Figure 5. Visualisation of group pairs and the analysis of significant differences

A. Differences in the mean number of oocyte count in the vitellogenic stage between the three group pairs. The mean pair differences between genotypes CPMP-CPCP ($P < 0.001$) and MPMP-CPMP ($P < 0.001$) are significant; but MPMP-CPCP is not ($P > 0.5$). **B.** Differences in the mean number of oocyte count in the choriogenic stage between the three group pairs. The mean pair differences between genotypes CPMP-CPCP ($P < 0.05$) and MPMP-CPCP ($P < 0.05$) are significant; but MPMP-CPCP is not ($P > 0.9$). In this visual representation significant differences are those that do not cross 0.

H. cydno genome and annotation transfer

The final *H. cydno* guided assembly has 58 scaffolds; with 22 191 fragments used for the assembly. The sum of length of the fragments used in the assembly is 261 056 210 bp. There are 13 774 unplaced fragments (17 915 151 bp, 6.63% total) and there are 7843 935 (3%) Ns introduced. The assembly N50 is 13 724 118 (*H. cydno* reference fasta file; ordering information and unplaced scaffolds available from https://www.dropbox.com/sh/5krc7kn3u0oviwj/AADHTlQsoxQCnqZnivatNdRba?dl=0).

The *H. melpomene* annotation has 193 665 elements. Of these 173 832 were transferred and 22 776 parts of elements (i.e. exons, tRNA) could not be transferred. There are 19 609 gene models in *H. melpomene* and of these, 17 478 were transferred to *H. cydno*. There are 3 784 exons that were not transferred from partial CDS matches; and 2 539 gene models not transferred. embl *H. cydno* annotation file was converted to gff (*H. cydno* gff file available from https://www.dropbox.com/sh/5krc7kn3u0oviwj/AADHTlQsoxQCnqZnivatNdRba?dl=0, Figure 8).

Sequencing, read mapping and counting feature abundance

Of 30 samples sequenced for this project, 13 are *H. melpomene*, 7 are *H. cydno* and 10 are *H. cydno* x *H. melpomene*. They have a median total number of reads of 34.93 M (min. 24.51M; max. 44.22 M) (Supplementary Table S1). *H. cydno* and *H. melpomene* had their most recent common ancestor 1.5 million years ago and their absolute divergence is roughly of 3% (dxy ~ 0.03) (Kozak *et al.*, 2015; Martin *et al.*, 2016; details in the Introduction). The most common source of error when measuring gene expression is miscalculation due to poor mapping reads. With an absolute divergence of 3% this was likely to be an issue when mapping the *H. cydno*

samples to the *H. melpomene* reference genome (Hmel2, Davey *et al.,* 2016). In an attempt to reduce this source of bias I generated a new *H. cydno* genome and annotation. After generating such reference genome and annotation, I performed tests with a subset of the mRNAseq samples to assess whether indeed the *H. cydno* genome was more appropriate than the *H. melpomene* genome as a reference for the *H. cydno* samples. Moreover, I wanted to access how do the new *H. cydno* assembly and the reference *H. melpomene* assembly compare to *de novo* assemblies. I also wanted to calculate how the *H. cydno* X *H. melpomene* hybrid samples scored when mapped to the two parents genomes. I observe the best overall alignment rate of mRNA read pairs when: 1) the *H. cydno* samples are mapped to the *H. cydno* genome (Cydno_paternal_ragout) and when, 2) the *H. melpomene* samples are mapped to the *H. melpomene* genome (Hmel2_ordered) (Figure 6). Hybrid sample overall alignment rates are lower than those observed for *H. melpomene* samples mapped to the *H. melpomene* genome; and *H. cydno* samples mapped to the *H. cydno* genome; but higher than those I obtain when I map the *H. melpomene* samples to the *H. cydno* genome and *H. cydno* samples to the *H. melpomene* genome.

Using the *H. cydno* genome assembly (Cydno_paternal_ragout) results in higher overall mapping percentages than just using the trio-sga assembly (Cydno_paternal_milan). As expected, mapping to the *H. melpomene* assembly also results in higher overall mapping percentages than using the trio-sga assembly (Melpomene_paternal_milan) (Figure 6). I achieved the greatest percentages of reads mapping when I used the genome that corresponded to the species of the sample and HISAT2 default options. Even though overall mapping percentage is increased by relaxing the mapping parameters, multiple mappings also increased. DESeq2 assigns multi-mapping reads randomly between the multiple mapping sites. Increasing the overall mapping percentage due to multi-mapping reads is therefore not helpful (Figure 6).

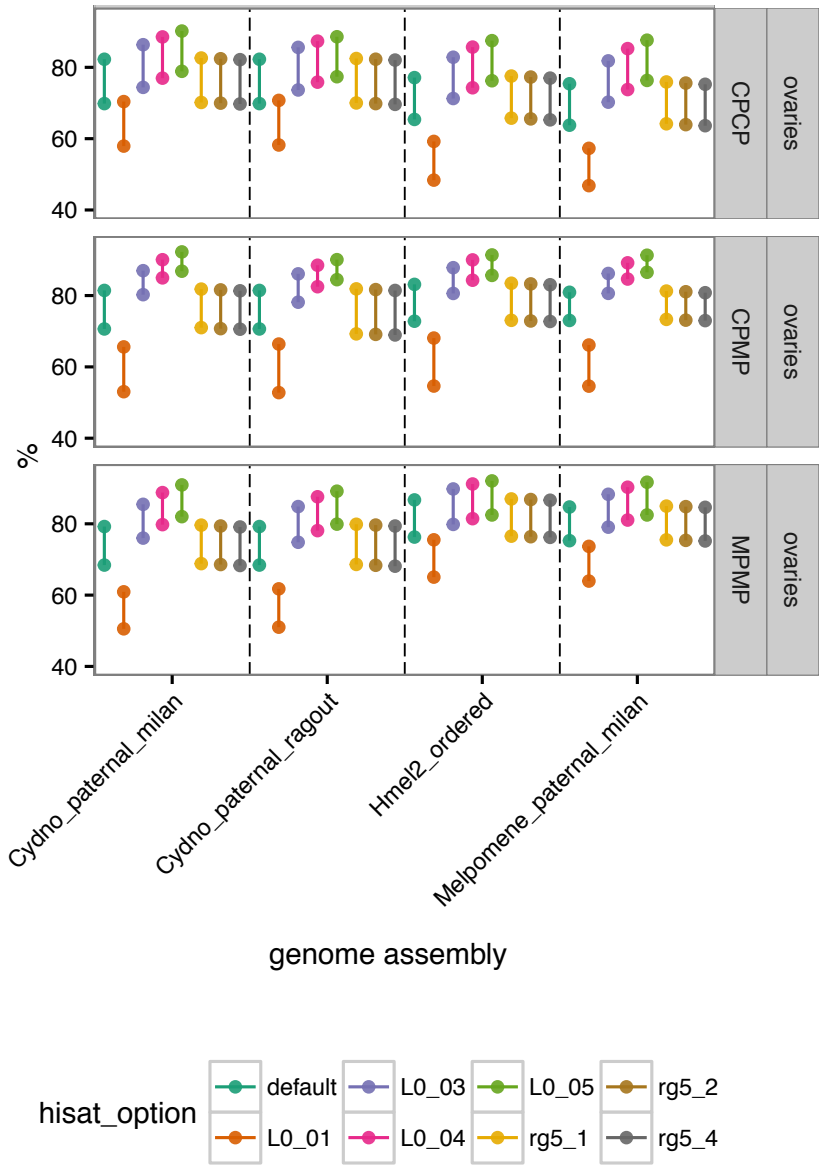


Figure 6. Samples with the highest and the lowest overall mRNA reads alignment percentage from the three different samples groups to the four different genome assemblies

Each point represents a sample, and there are two samples per group: 1) the sample from the whole sequenced set with the highest overall mapping percentage; 2) and the one with the lowest mapping

percentage. Overall mapping percentages are plotted for each one of the two samples of each group: 1) *H. cydno* ovaries – top panel, CPCP; 2) *H. cydno* x *H. melpomene* ovaries – middle panel, CPMP; 3) *H. melpomene* ovaries – bottom panel, MPMP. Overall mapping percentages (%) for these samples are shown in the y-axis. The x-axis has the different genome assemblies tested: 1) Cydno_paternal_milan: trio-sga *H. cydno* assembly; 2) Cydno_paternal_ragout: Ragout *H. cydno* guided assembly; 3) Hmel2_ordered: *H. melpomene* assembly; 4) Melpomene_paternal_milan: trio-sga *H. melpomene* assembly. The different colours represent the different mapping parameters used. L0_x command line option changes the maximum number of ambiguous character allowed in the read as a function of the read length. rgx_x command line option sets the read gap open and extends penalties.

A feature is an interval (i.e. range of positions) in a genome. In this analysis I used gene models as the features of interest. After aligning the sequencing reads with the list of genes for each annotation I counted how many read pairs mapped to each gene. Perhaps unsurprisingly, there are also more features counted using when I use the *H. cydno* annotation to count genes from *H. cydno* samples; and *H. melpomene* annotation (v2, Chapter 2, “Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*”) to count genes from *H. melpomene* samples. *H. cydno* x *H. melpomene* samples have more variance on those read pairs associated to no feature than both *H. cydno* and *H. melpomene*. There are also a higher number of read pairs without a feature when the F1 samples are counted against the *H. melpomene* annotation (Figure 7).

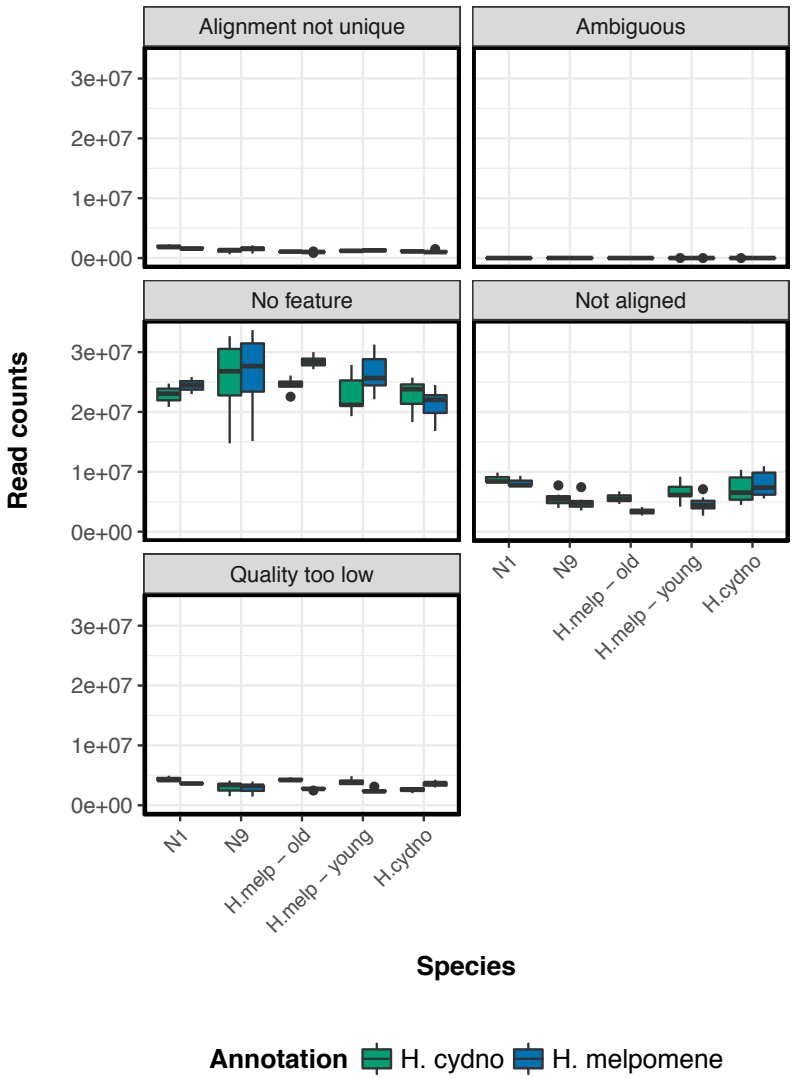


Figure 7. Total number of read pairs mapping to genic features defined in the *H. melpomene* (Hmel2.1) and *H. cydno* (RATT transfer) annotation

All samples were mapped to the *H. cydno* Ragout reference and to the *H. melpomene* reference genome. Then total number of read pairs mapping to genic features defined in the *H. melpomene* (Hmel2.1) and *H. cydno* (RATT transfer) annotation were counted. Box-and-whisker plots representing the number of read pairs mapping: 1) Alignment not unique – read pairs with more than one reported alignment; 2) Ambiguous – read pairs which could have been assigned to more than

one feature and so are not counted; 3) No feature – read pairs which could not be assigned to any feature; 4) Not aligned – read pairs in the SAM file without alignment to the reference genome; 5) Quality too low – read pairs with an alignment quality lower than 10 (default value). Number of read pairs for the two hybrid female broods (N9 and N1); young and old *H. melpomene* and young *H. cydno* represented separately.

Gene expression clusters individuals by group when mapping to either reference genome/annotation

Gene expression profiles cluster the samples by group: *H. melpomene*, *H. cydno* and *H. cydno* x *H. melpomene*. When the read pairs are mapped to the *H. melpomene* genome and genic features are counted using the *H. melpomene* annotation (v2), 43% of the total variance is explained by the two first principal components. PC1 separates samples by species and explains 24% of the variance. *H. melpomene* samples when mapped to the *H. melpomene* genome and annotation show more diversity than *H. cydno* samples which form a very tight cluster (Figure 8).

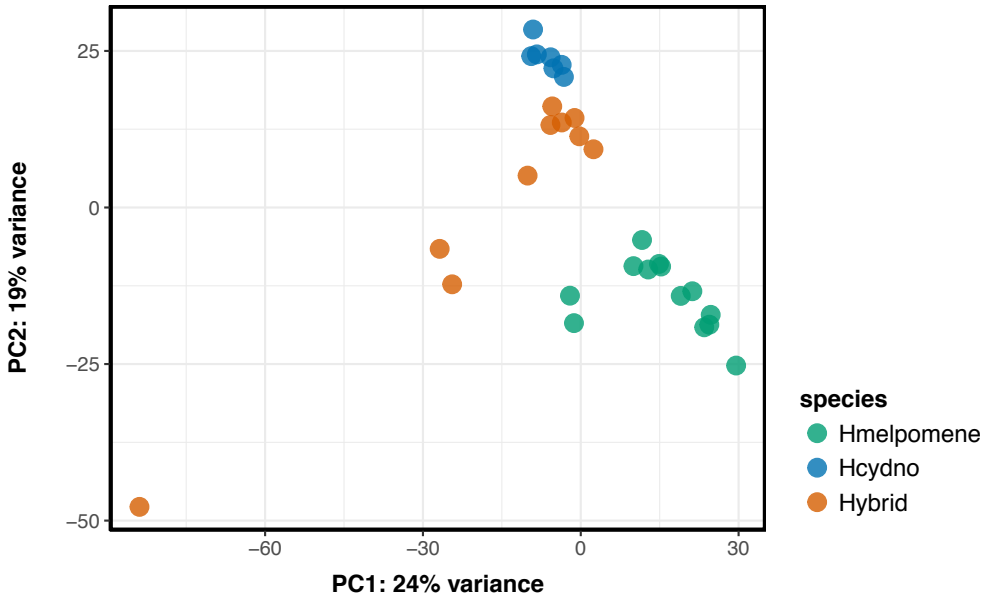


Figure 8. Principal component analysis of gene expression profiles for the 30 ovary samples quantified mapping to the *H. melpomene* reference genome and annotation

PCA of ovary tissue transformed gene expression count data (log2, DESeq2, rlog(blind=FALSE)). rlog transformed data minimizes differences between samples for rows with small counts and normalises with respect to library size.

When the read pairs are mapped to the *H. cydno* genome and genic features are counted using the *H. cydno* annotation, 42% of the total variance is explained by the two first principal components. PC1 explains 25% of the variance. *H. cydno* x *H. melpomene* samples when mapped to the *H. cydno* genome and annotation separate along PC2 by brood. *H. cydno* samples form a less tight cluster than in when the samples are mapped to *H. melpomene* genome/annotation. Hybrid samples only separate from *H. cydno* samples in PC2 (Figure 9).

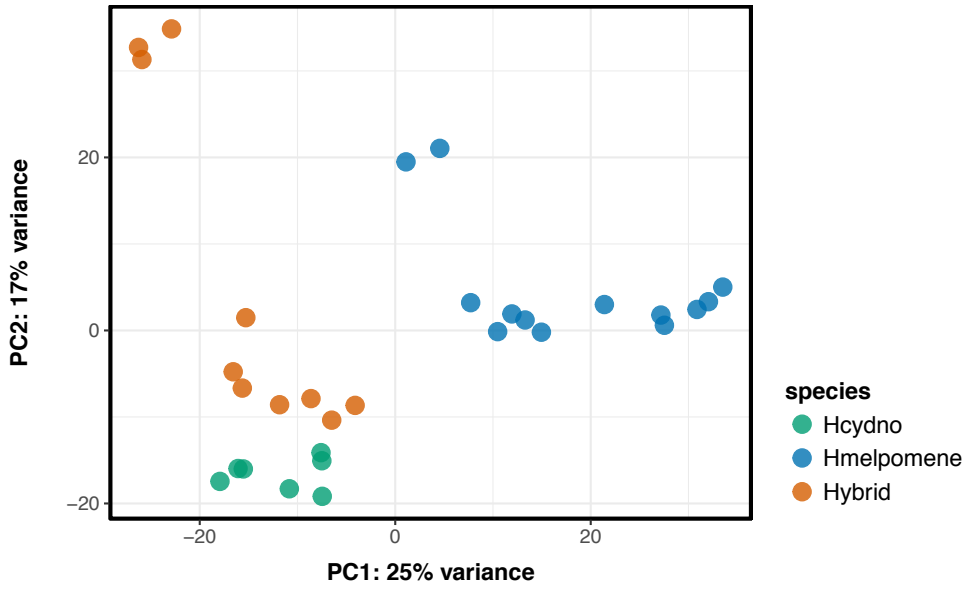


Figure 9. Principal component analysis of gene expression profiles for the 30 ovary samples quantified mapping to the *H. cydno* reference genome and annotation

PCA of ovary tissue transformed gene expression count data (log2, DESeq2, rlog(blind=FALSE)). rlog transformed data minimizes differences between samples for rows with small counts and normalises with respect to library size.

Differentially expressed genes between *H. cydno, H. melpomene* and the hybrids

Overall there are a total of 498 genes differentially expressed between the three groups when analysis used one or both annotations. 90 of these did not get transferred to *H. cydno* annotation so their expression pattern was only quantified when the analysis was carried out using the *H. melpomene* reference genome/annotation. There are 104 genes that are present in both annotations but were only differentially expressed when the analysis was

done using the *H. cydno* annotation. Finally, there are 246 genes that are also present in both annotations but are only classified as differentially expressed using the *H. melpomene* annotation (Figure 10).

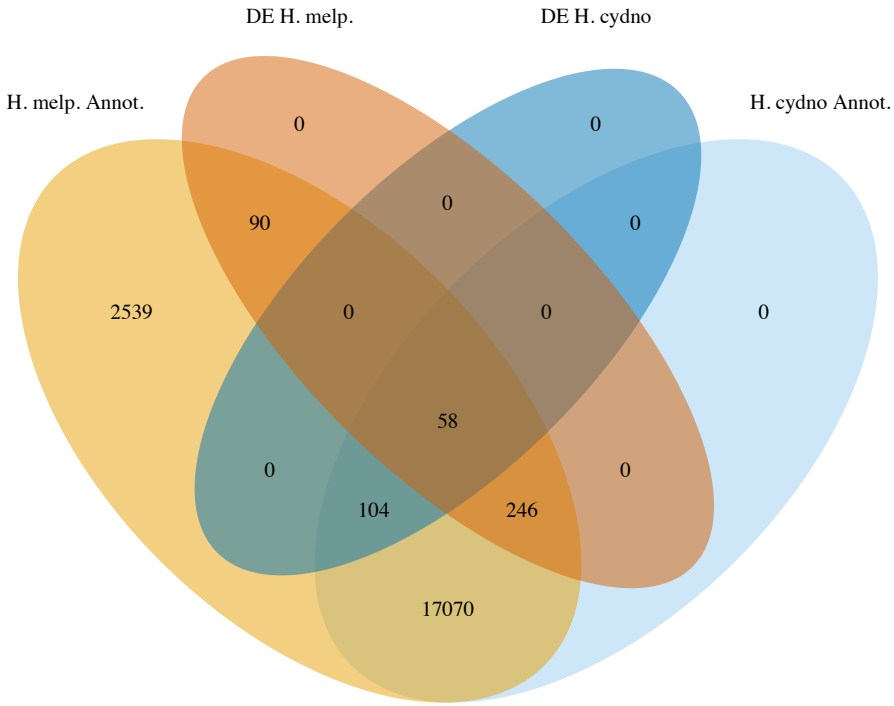


Figure 10. Venn diagram summary of the genes classified as differentially expressed between the three groups

The *H. melpomene* annotation has a total of 19 609 predicted genes (H. melp. Annot.). Of these 17 478 were transferred to *H. cydno* (H. cydno Annot.). There are 90 genes that did not get transferred to *H. cydno* and are classified as differentially expressed when the analysis is done using the *H. melpomene* reference genome and transcriptome. 58 genes are classified as differentially expressed using both annotations; 104 just with the *H. cydno* annotation and 246 (DE H. cydno, Supplementary Figure S2) just with the *H. melpomene* annotation (DE H. melp, Supplementary Figure S1).

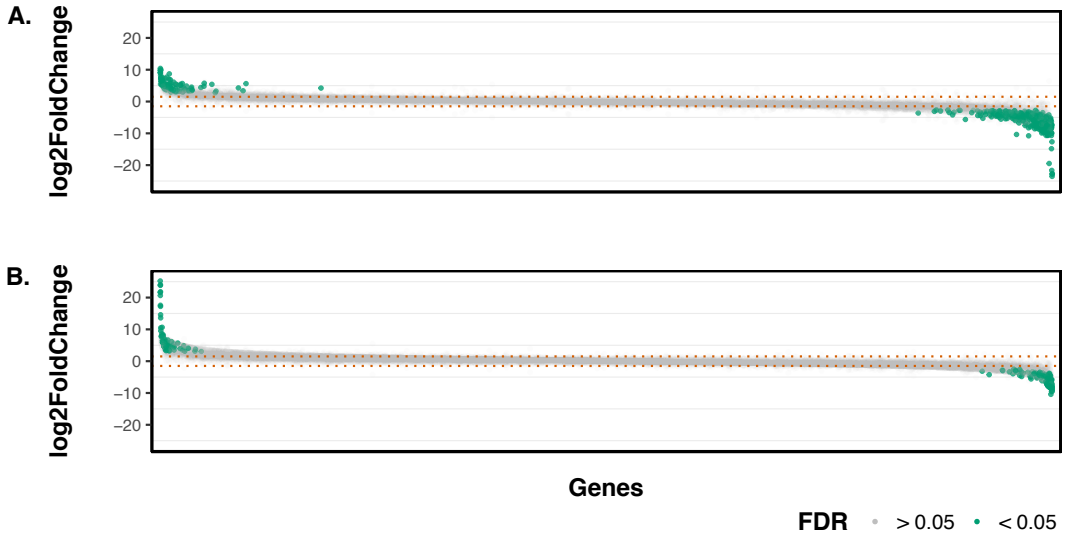


Figure 11. Differential gene expression between the *H. melpomene*, the *H. cydno* and the hybrids

A positive log2 fold change means that the number of transcripts is higher in the hybrids and/or *H. cydno* with the *H. melpomene* samples as the baseline (**A**); the hybrids and/or *H. melpomene* using the *H. cydno* samples as a baseline (**B**); and vice-versa. y axis: log2 fold change; x axis: individual genes ordered by log2 fold change value. Significant log2 fold changes in transcript abundance between the groups are represented in green (FDR<0.05). Dotted lines indicate 1.5 log2 fold change threshold.

By performing the differential expression analysis with the *H. melpomene* reference genome/annotation and the *H. cydno* reference genome/annotation the set of differential expressed genes differs. This is not surprising if we consider how mapping and counting success varies when *H. cydno* read pairs are mapped and counted against the *H. melpomene* reference genome/annotation and vice-versa. For the rest of the analysis I will,

therefore, consider all genes with differential expression in both analyses (498 genes in total). I also compared genome-wide trends for the genes found with both reference genomes/annotations separately (Supplementary Figures S1, S2 and S3) and found similar results except where stated below.

The ends of chromosomes are enriched with differentially expressed genes

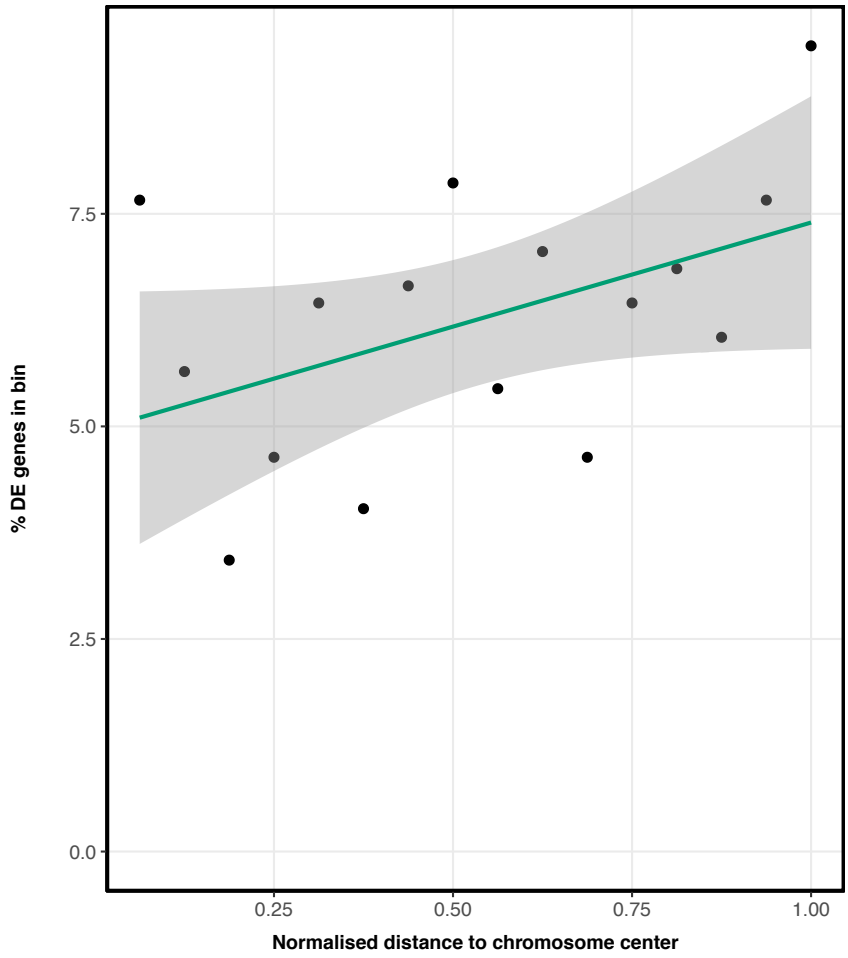


Figure 12. Distribution of differentially expressed genes along chromosome position

Total differentially expressed genes grouped into 16 bins along their normalised distance from the centre of the chromosome. In the x-axis, 0 is the normalised chromosome centre, 1 is the normalised chromosome end. Linear regression line fit to the correlation between total number of differentially expressed genes per bin and distance to chromosome centre.

The total of 498 differentially expressed genes tend to map towards the end of chromosomes (Adj R2 = 0.157, Intercept =-0.012, Slope = 8.693) and there is a depletion of differentially expressed genes in towards the centres of chromosomes (Figure 12; Supplementary Figure S3).

Also, differentially expressed genes are not equally distributed among the different chromosomes (Figure 13A) although there is not a strong correlation between chromosome size and total number of differentially expressed genes (Adj R2 = 0.022, Intercept =-17.5, Slope = 4.23, Figure 13B). Chromosome 1, with 258 differentially expressed genes and a total length of ~17Mb, is the largest chromosome in the genome and also the one with the most differentially expressed genes. However, chromosome 2, with 93 differentially expressed genes and a total length of ~9Mb, is the smallest chromosome in the genome but has the second largest number of differentially expressed genes identified (Figure 13B). Significant expression differences were found on 62.4% of the scaffolds. 8.5% of the scaffolds have only one differential expressed gene and only 0.8% have 10 or more differentially expressed genes. There are 5 scaffolds with more than 14 differentially expressed genes. These 5 scaffolds are located in 5 different chromosomes: chromosome 2 (Hmel202006, 20 genes); chromosome 17 (Hmel217020, 17 genes); chromosome 9 (Hmel209007, 16 genes); chromosome 20 (Hmel220005, 14 genes) and chromosome 7 (Hmel207002, 14 genes) (Figure 13C).

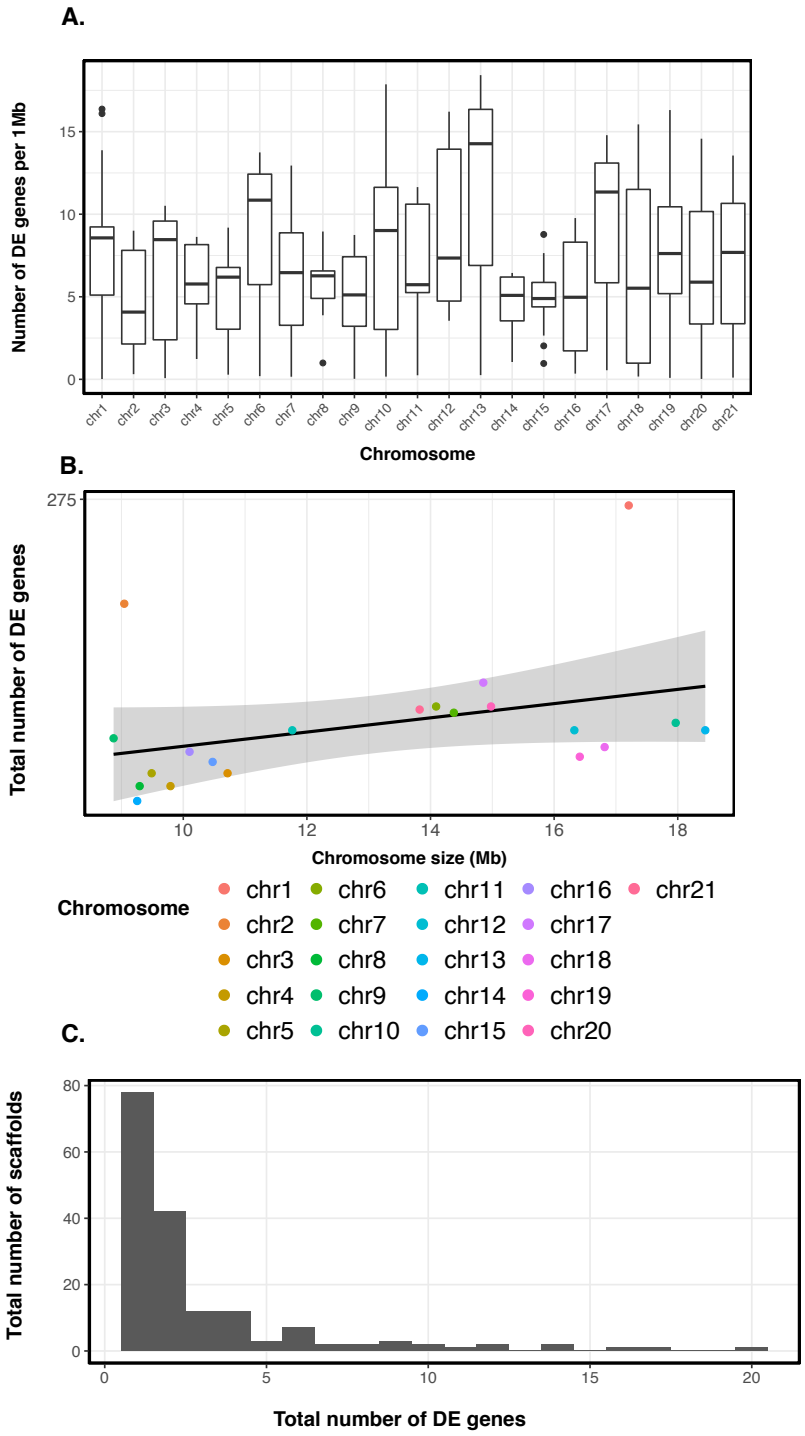


Figure 13. Genome-wide distribution of differentially expressed genes

A. Box-and-whisker plots displaying the 498 differentially expressed genes per 1Mb-window for each chromosome. **B.** Total number of differentially expressed genes is weakly correlated to chromosome size. Each point represents one chromosome. Chromosome 1 and 2 have a total number of differentially expressed greater than other chromosomes.

Differentially expressed genes and the predicted biological processes, cellular component, molecular function and protein class they are associated with

The differential expressed genes identified are not equally distributed along the genome and they are predicted to be involved in several different biological processes, cellular components, have different molecular functions, and belong to several different protein classes. Of the 497 PANTHER IDs, 178 could not be mapped against the *D. melanogaster* genome in order to do the enrichment analysis and 114 had multiple mapping information. 206 sequences were successfully assigned a biological process, molecular function, cellular component or protein class. Some sequences had predicted functions in more than one category (Table 5, Figure 14).

Identifier	Category	Description
GO:0071840	Biological process	Cellular component organization or biogenesis
GO:0009987	Biological process	Cellular process
GO:0051179	Biological process	Localization
GO:0065007	Biological process	Biological regulation
GO:0000003	Biological process	Reproduction
GO:0050896	Biological process	Response to stimulus
GO:0032502	Biological process	Developmental process
GO:0032501	Biological process	Multicellular organismal process
GO:0008152	Biological process	Metabolic process

GO:0016020	Cellular component	Membrane
GO:0032991	Cellular component	Macromolecular complex
GO:0044464	Cellular component	Cell part
GO:0043226	Cellular component	Organelle
GO:0005576	Cellular component	Extracellular region
GO:0045182	Molecular function	Translation regulator activity
GO:0005488	Molecular function	Binding
GO:0004872	Molecular function	Receptor activity
GO:0005198	Molecular function	Structural molecule activity
GO:0004871	Molecular function	Signal transducer activity
GO:0003824	Molecular function	Catalytic activity
GO:0016209	Molecular function	Antioxidant activity
GO:0005215	Molecular function	Transporter activity
PC00102	Protein class	Extracellular matrix protein
PC00085	Protein class	Cytoskeletal protein
PC00227	Protein class	Transporter
PC00220	Protein class	Transferase
PC00176	Protein class	Oxidoreductase
PC00144	Protein class	Lyase
PC00069	Protein class	Cell adhesion molecule
PC00142	Protein class	Ligase
PC00171	Protein class	Nucleic acid binding
PC00207	Protein class	Signaling molecule
PC00095	Protein class	Enzyme modulator
PC00060	Protein class	Calcium-binding protein
PC00090	Protein class	Defense/immunity protein
PC00121	Protein class	Hydrolase
PC00219	Protein class	Transfer/carrier protein
PC00150	Protein class	Membrane traffic protein
PC00218	Protein class	Transcription factor
PC00070	Protein class	Cell junction protein
PC00135	Protein class	Isomerase
PC00197	Protein class	Receptor

Table 5. List of predicted biological processes, cellular components, molecular functions and protein classes for the whole set of differential expressed genes

Reference sequences of the differentially expressed genes were probed against the InterPro database.

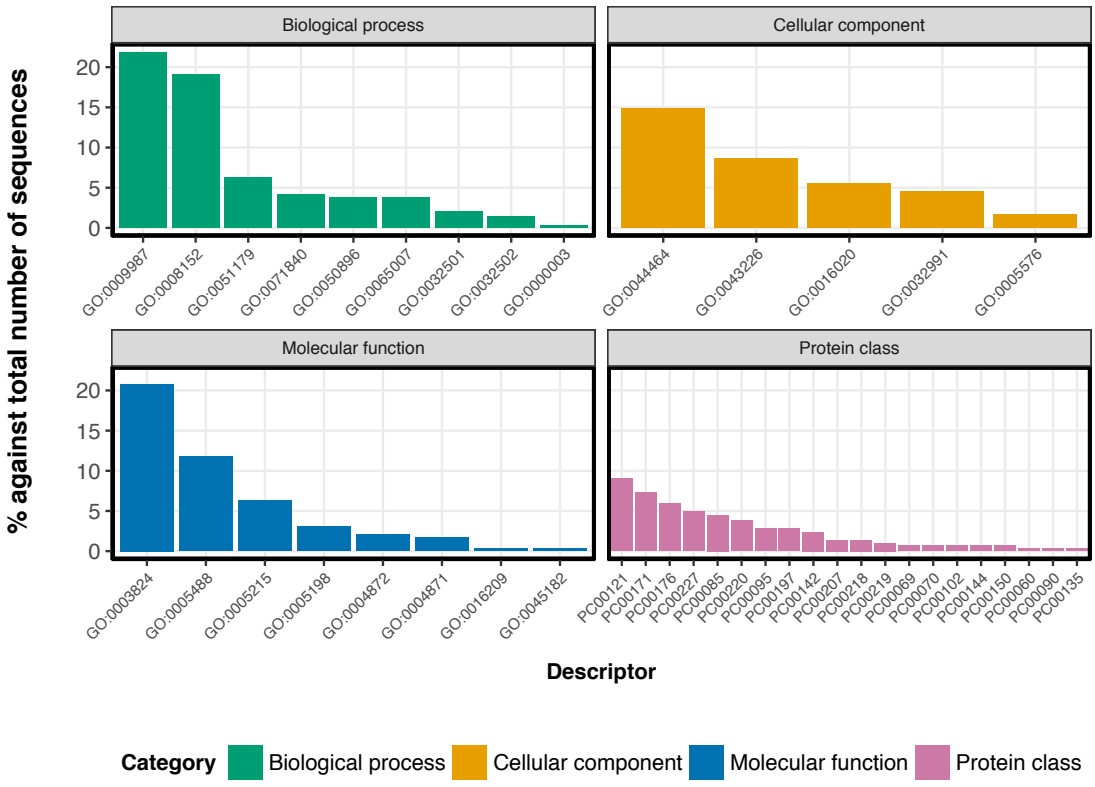


Figure 14. Distribution of predicted biological processes, cellular components, molecular functions and protein classes for the whole set of differential expressed genes

Reference sequences of the differentially expressed genes were probed against the InterPro database. x-axis term key from Table 5.

I performed enrichment analysis to quantify whether any biological functions were over-represented among the whole set of differentially expressed genes identified using both the *H. cydno* and *H. melpomene* reference genomes/annotations. The total numbers in any category were small and so there were no gene ontologies significantly enriched following Bonferroni correction for multiple testing ($P > 0.45$).

Twelve differentially expressed genes overlap the sterility QTL in chromosome 21

Using backcrosses of *H. cydno* females with *H. cydno* x *H. melpomene* males Merrill *at al.* (unpublished) has identified a QTL on the Z chromosome that is significantly associated with hybrid sterility. This sterility QTL spans 470 genes and 14 scaffolds (61 57 374 bp) with the peak in scaffold Hmel221012 at 1 912 456 bp. In total I found 31 differentially expressed genes in the sex chromosome and 12 overlap the identified sterility QTL (Figure 15).

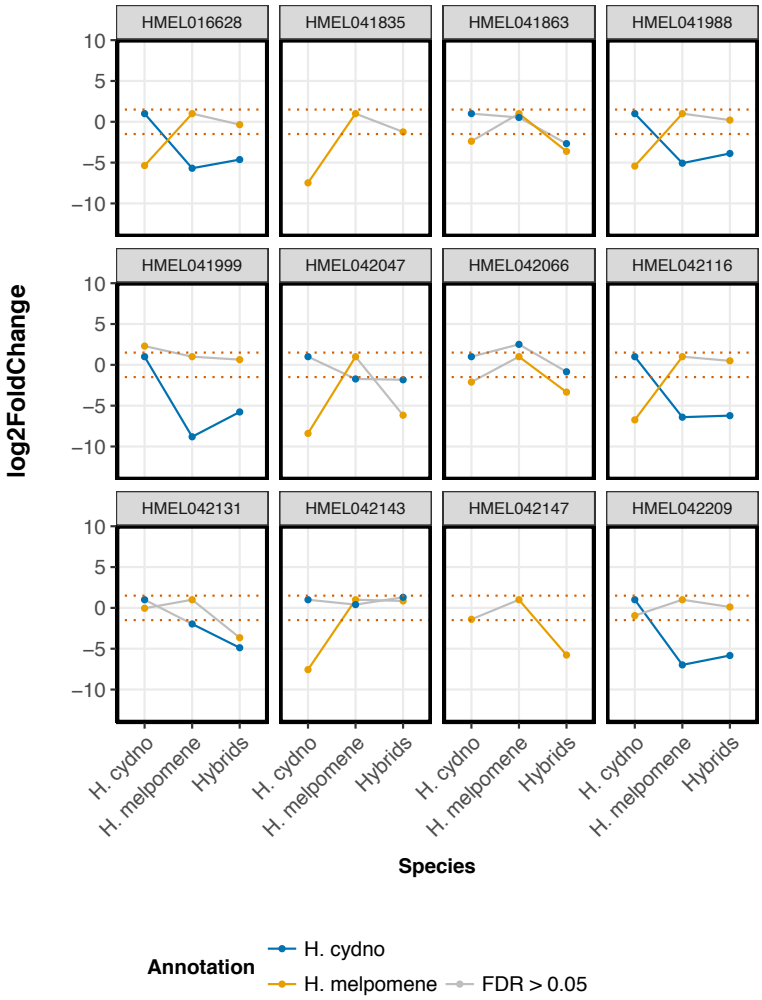


Figure 15. Patterns of expression for the differential expressed genes that overlap the sterility QTL

Genes differentially expressed identified with the analysis performed on the *H. cydno* (blue) and the *H. melpomene* (yellow) reference genome and annotation. Points represent a gene in each of the three different sample groups. Gene expression for each gene calculated with *H. cydno* and the *H. melpomene* samples as the baseline. Gene expression values of the three different groups (*H. cydno*, *H. melpomene* and hybrids) is linked by blue, yellow or grey lines. Blue/yellow lines represent significant results (FDR < 0.05 and log2 fold

change > |1.5|), grey lines represent non-significant results (FDR > 0.05 and/or log2 fold change < |1.5|). Dotted red lines delineated the |1.5| log2 fold change significance threshold. Genes HMEL041835 and HMEL042147 did not get transferred do the *H. cydno* annotation.

I selected the genes that exhibited the same expression pattern in the *H. cydno* and *H. melpomene* but that were either up- or down-regulated in the hybrids. These genes are HMEL041863 (Hmel221009: 105514-129779); HMEL042066 (Hmel221012: 2108292-2108576); HMEL042131 (Hmel221013: 40852-41133) and HMEL042147 (Hmel221014: 2795-3501). There are no 1-1 orthologues between *H. melpomene* and *H. erato* and the only gene with a predicted function is HMEL042147, which is homologous to Lethal (3) malignant brain tumour (L(3)mbt) in *Drosophila*, is a tumour suppressor protein regulating proliferation in the brain particularly the optic lobes. Within the genes that have the same expression pattern in the *H. cydno* and *H. melpomene* but differ in the hybrids, HMEL042066 is the closest to the sterility QTL peak (~126 Kb apart). Expression patterns in *H. melpomene* and *H. cydno* for genes HMEL016628, HMEL041988 and HMEL042116 is not consistent if different reference genomes are used. For all 3 genes, if expression is quantified using the *H. cydno* genome the *H. melpomene* samples are classified down-regulated. However, if expression is quantified using the *H. melpomene* genome than *H. cydno* samples are classified as down-regulated. Regardless of which annotation is used Hybrid samples are always at the same level as the reference. This inconsistency is likely to reflect sequence divergence between the gene's *H. cydno* and *H. melpomene* reference and it may be interesting to focus on such genes as strong candidates of BDMIs.

Loci with no gene flow between *H. cydno* and *H. melpomene* are over-represented in the differentially expressed dataset

First I compared overlap between differentially expressed genes and loci considered to be putative species barriers in a demographic model fitted to population genomic data. Camille Roux computed posterior probabilities for two demographic models of isolation and gene flow fitted to whole genome data from *H. melpomene rosina* and *H. cydno chioneus* using an Approximate Bayesian computation (ABC) approach. Each gene in the genome was considered as a separate locus and two models compared for each gene. The model assumes independence with partial linkage within loci. Model 1 is a migration model in which the locus is free to introgress, while model 2 is a non-migration model in which the locus is linked to a barrier. The scenarios were simulated 6 million times to estimate a posterior probability of fit to either model. For example, a pattern of polymorphism and divergence observed at a protein coding locus is better explained by a model with no migration and can be a barrier region.
In the whole genome there are 9396 loci scored as allowing free gene flow and 399 loci with no gene flow. Of these 350 and 23 are differentially expressed genes respectively, with more loci showing differential expression than expected at random (Pearson's $X^2 = 9.11$, $P < 0.05$).

These 23 loci of no gene flow that show differential expression in my dataset span 19 genes. Two of these genes have predicted biological functions of development and reproduction and map to chromosome 7 and to chromosome 20. The other 17 genes map to 11 different chromosomes. None of these 19 genes overlap duplications identified between *H. cydno* and *H. melpomene* (Pinharanda *et al.*, 2017) or the sterility QTL identified in chromosome 21 (Figure 16, Table 6).

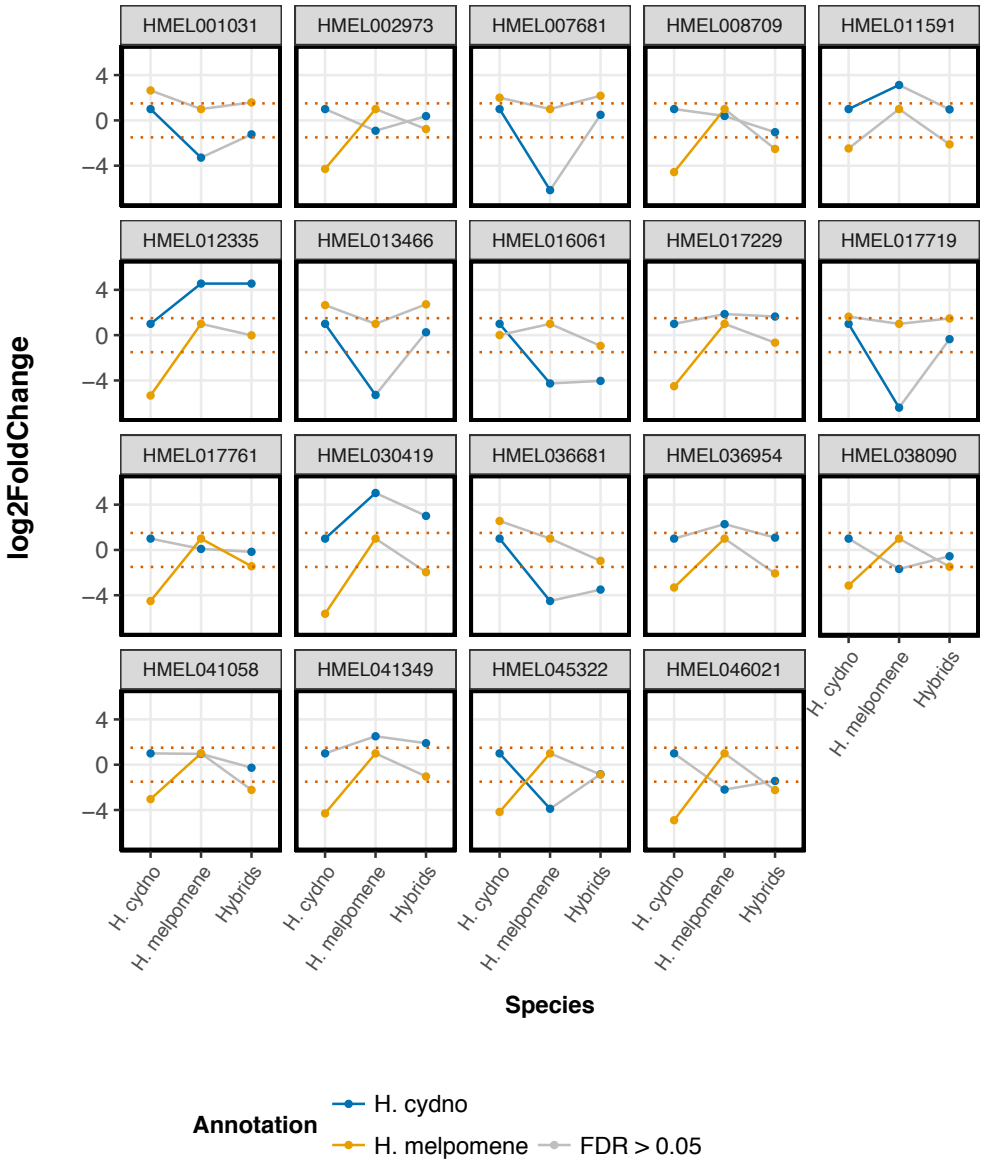


Figure 16. Patterns of expression for the differential expressed genes that loci lacking gene flow between *H. cydno* and *H. melpomene*

Genes differentially expressed identified with the analysis performed on the *H. cydno* (blue) and the *H. melpomene* (yellow) reference genome and annotation. Points represent a gene in each of the three different sample groups. Gene expression for each gene calculated with *H. cydno* and the *H. melpomene* samples as the baseline. Gene

expression values of the three different groups (*H. cydno*, *H. melpomene* and hybrids) is linked by blue, yellow or grey lines. Blue/yellow lines represent significant results (FDR < 0.05 and log2 fold change > |1.5|), grey lines represent non-significant results (FDR > 0.05 and/or log2 fold change < |1.5|). Dotted red lines delineated the |1.5| log2 fold change significance threshold.

Gene	Chr	Scaffold	Start	End	Ortho.	α
HMEL002973	chr1	Hmel201002	2045576	2050494	3012	0
HMEL017229	chr1	Hmel201009	2691600	2697252	3164	NA
HMEL030419	chr1	Hmel201009	2701980	2704285	NA	NA
HMEL036681	chr2	Hmel202004	1155648	1158588	3535	NA
HMEL036954	chr2	Hmel202006	1994862	2001702	NA	NA
HMEL045322	chr6	Hmel206019	184458	189406	4850	NA
HMEL016061	chr7	Hmel207002	23341	38406	NA	NA
HMEL046021	chr7	Hmel207002	4786068	4787106	5193	0
HMEL011591	chr9	Hmel209007	5170835	5172279	NA	NA
HMEL012335	chr12	Hmel212013	3320402	3327743	7539	1
HMEL007681	chr14	Hmel214024	265221	273936	8250	NA
HMEL008709	chr15	Hmel215035	169591	176717	8579	0
HMEL013466	chr17	Hmel217020	2507983	2515279	NA	NA
HMEL017719	chr17	Hmel217001	1089513	1095776	8940	NA
HMEL017761	chr17	Hmel217004	4314606	4317581	9129	0.52
HMEL038090	chr17	Hmel217004	795261	800046	8996	0.52
HMEL001031	chr18	Hmel218003	614600	619104	9415	0
HMEL041349	chr20	Hmel220012	370319	376337	10617	0.07
HMEL041058	chr20	Hmel220005	2285054	2287739	10433	NA

Table 6. Differentially expressed genes that overlap regions of the genome with putatively no gene flow between *H. cydno* and *H. melpomene*

List of differentially expressed genes that overlap regions of the genome with putatively no gene flow between *H. cydno* and *H. melpomene*. Ortho – Orthogroups from Chaper 2 between *H. melpomene* and *H. erato*. Orthogroups start with OG00 followed by number in the Ortho. column. α – rate of adaptive evolution calculated in Chapter 2, "Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*". If orthogroup was identified but there is no α estimated, there were too many undetermined characters in the sequence to estimate parameters.

Overall, loci with putatively no gene flow between *H. cydno* and *H. melpomene* map in excess to the Z chromosome. The distribution of the number of loci supported by a model of no gene flow in sex-linked loci clearly shows that there are significantly more loci predicted to have no gene flow in the Z than expected (10 000 times without replacement; $P < 0.001$) (Figure 17). Interestingly, the sterility QTL identified by Richard Merrill also maps to the sex chromosome and the region overlaps significantly with putative no gene flow loci than expected by change (Pearson's $X^2 = 40.879$, $P < 0.001$).

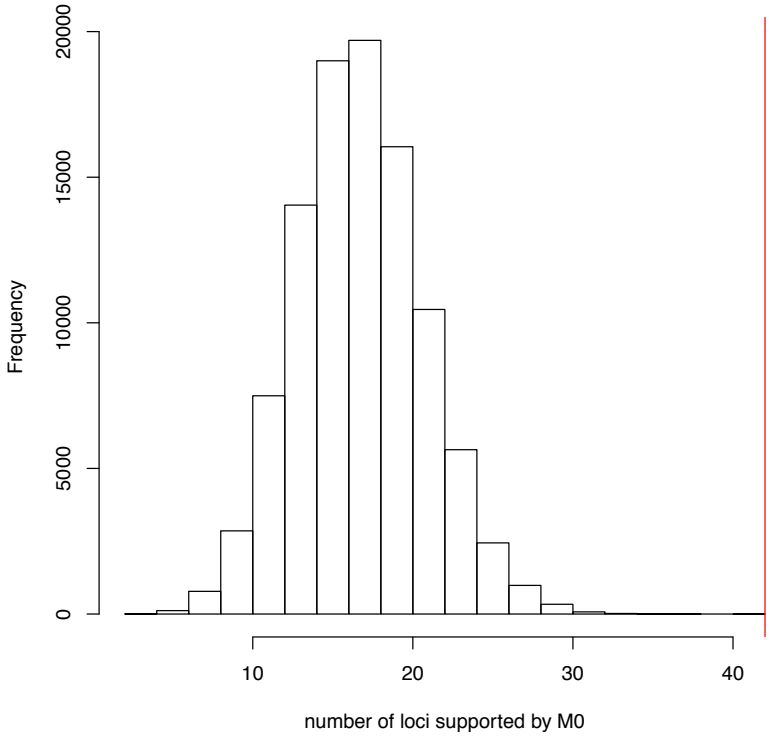


Figure 17. Distribution of the number of loci supported by a model of no gene flow between *H. cydno* and *H. melpomene*

Distribution obtained by resampling without replacement 10 000 times sex-linked loci among all loci. The number of sex-linked loci supporting a model of no gene flow observed in the real data indicated in red.

H. cydno specific reads are over-represented in the expressed transcripts of *H. melpomene* X *H. cydno* samples

All the samples analysed had a higher proportion of *H. cydno* mother specific reads than *H. melpomene* father specific reads (Figure 18A, Table 7). Moreover, for all the samples, there were also more of the *H. cydno* specific reads mapping to genic features than *H. melpomene* specific reads (Figure 18B, Table 7).

Sample	Brood	Species	RP no mismatches	RP specific	RP feature
AP52	N9	H. melp.	456850	196434	34639
AP53	N9	H. melp.	631034	284236	49304
AP54	N9	H. melp.	584150	267890	55171
AP58	N9	H. melp.	768216	360882	62145
AP65	N9	H. melp.	736656	347052	59514
AP66	N9	H. melp.	546866	216138	49811
AP70	N9	H. melp.	306758	128712	21754
AP38	N1	H. melp.	968596	304690	125114
AP39	N1	H. melp.	626438	228710	72298
AP41	N1	H. melp.	789236	332714	131754
AP52	N9	H. cyd.	456850	260416	73743
AP53	N9	H. cyd.	631034	346798	118334
AP54	N9	H. cyd.	584150	316260	104674
AP58	N9	H. cyd.	768216	407334	132819
AP65	N9	H. cyd.	736656	389604	119792
AP66	N9	H. cyd.	546866	330728	89589
AP70	N9	H. cyd.	306758	178046	54965
AP38	N1	H. cyd.	968596	663906	273641
AP39	N1	H. cyd.	626438	397728	131539
AP41	N1	H. cyd.	789236	456522	146416

Table 7. Summary of read filtering totals for the *H. cydno* X *H. melpomene* samples using each parent's alternative reference

Brood names are given per sample. RP – read-pair. For each sample: 1) the total number of reads with no mismatches when mapped to either parent's reference (RP no mismatches); 2) the total number of species specific read pairs (RP specific); and the 3) total number of read pairs mapping to genic features (RP feature) are shown.

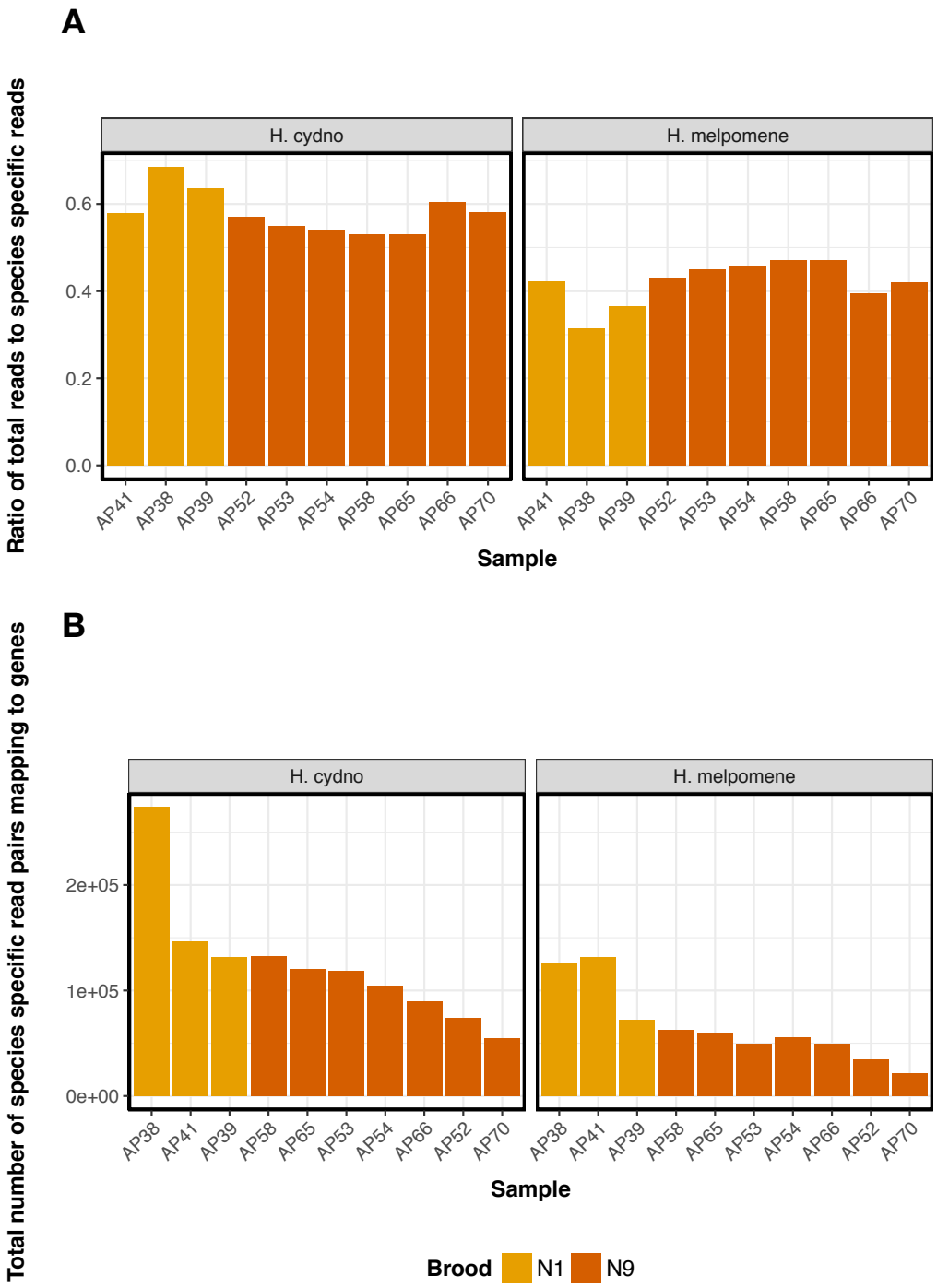


Figure 18. *H. cydno* specific reads are present in hybrid transcripts and map to genic features more often than *H. melpomene* reads

Facets separate the samples by whether they were mapped to the *H. cydno* or to the *H. melpomene* genome and annotation reference. Colours correspond to the broods from which the samples were collected (N1 or N9). **Figure 18A** Ratio of total filtered reads vs. total species specific reads. x-axis bars correspond to each sample; y-axis represents the ratio of total reads to species specific reads. **Figure 18B** Total number of species specific read pairs mapping to genic features. X-axis bars correspond to each sample; y-axis counts the total number of species specific read pairs mapping to genes.

cis-regulatory differences represent most of the expression differences between *H. cydno* and *H. melpomene*

cis-regulatory divergence is expected to result from changes in DNA sequences close to the affected gene and relative allelic expression in hybrids directly correlates to *cis*-regulatory activity. *cis*-regulatory divergence is more common between *H. cydno* and *H. melpomene* than *trans*-regulatory differences (Fisher's exact test $P < 0.05$, Figure 19).

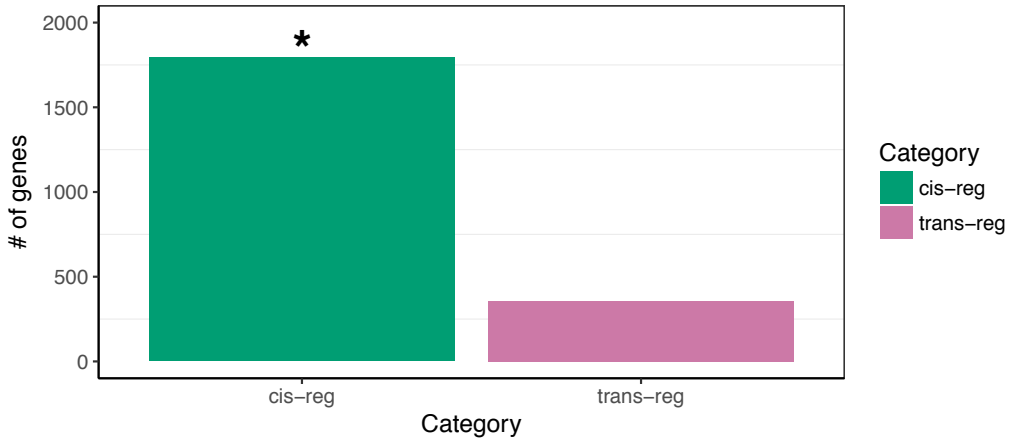


Figure 19. *cis*-regulatory differences represent most of the expression differences between *H. cydno* and *H. melpomene*.

cis-regulatory divergence is common between *H. cydno* and *H. melpomene* (Fisher's exact test $P < 0.05$). Significance represented by *. y-axis represents the number of genes and x-axis the divergence categories.

Discussion

Speciation is a complex process generally involving divergence along multiple phenotypic axes and involving the accumulation of many genetic changes. *H. cydno* and *H. melpomene* are hybridising species that show multiple pre- and post-zygotic barriers but also genome-wide signals of hybridization (Jiggins *et al.*, 2001; Naisbit *et al.*, 2002; Merrill *et al.*, 2012; 2013; Martin *et al.*, 2013). Recent work has focused on the role of wing patterns and behavioural change in speciation, but here I have conducted the first genome-scale study of hybrid breakdown in these species and the first attempt to identify F1 hybrid female sterility loci. I have identified genes distributed widely across the genome that show patterns of transgressive expression in hybrids consistent with a potential role in causing hybrid sterility.

In over 200 newly emerged dissected females, the ovaries of both species only contained pre-vitellogenic oocytes at the time of emergence (Dunlap-Pianka *et al.*, 1977). Vitellogenic oocytes were visible 3 to 4 hours after eclosion. In infertile F1 females there is a failure to complete the oocyte maturation process, suggesting that the sterility phenotype is likely to result from disruption of gene expression during the first few hours of adult life. The dissections, and subsequent gene expression analysis, reported in this study were preformed in young butterfly ovaries (~3h after eclosion) and so are likely to span the interval where oogenesis is disrupted in the *H. cydno* X *H. melpomene* female hybrid.

Hybrid sterility is a complex trait that likely results from disruption to gene expression caused by epistatic interactions between parental genomes. BDMIs arise when alleles that are neutral or beneficial in the parental genetic background lead to deleterious effects in a hybrid background. Fast evolving genes are more likely to be involved in incompatibility, and indeed most of the candidate differentially expressed genes did not have a 1 to 1 orthologue in *H. erato* (Chapter 2, "Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*"), consistent with rapid evolution (Tang and Presgraves, 2009). However it is important to note that, BDMIs do not necessarily have to show high divergence at the onset speciation (Nosil *et al.*, 2009). Overall, there were only 8 genes for which was possible to calculate the rate of adaptive evolution, and these do not have a high α (average α = 0.26). This number of genes is small but might suggest that the phenotype arose by drift rather than selection, perhaps subsequent to evolution of pre-zygotic reproductive barriers (Coyne, 1985; Rundle and Nosil, 2005). I have not considered mtDNA expression as mitochondrial scaffolds are not included in the primary *H. melpomene* scaffolds. As such, mito-nuclear interactions have not been considered. Expression of mtDNA may have strong trans-effects when divergent mitochondrial haplotypes are introgressed into the same nuclear background. However, mtDNA trans-effects are more likely to affect male gene expression and so less probable to have a correlation with hybrid female *H. cydno* x *H. melpomene* (Innocenti *et al.*, 2011; Camus *et al.,* 2015).

Two *rules of speciation* have been used to describe the genetic basis of post-zygotic isolation: Haldane's Rule and the large Z-effect. One of these, Haldane's Rule, is observed between these *H. cydno* and *H. melpomene* crosses. The other, the large Z-effect, is based on the observation that sex chromosomes have a significantly greater impact on hybrid fitness compared to autosomes (Coyne and Orr, 1998). In *Drosophila*, for example, it has been shown that there is a higher density of male sterility factors in the X chromosome (Masly and Presgraves, 2007). Here, there is mixed evidence for

a large Z-effect in Heliconius. Previous work, and the large-effect sterility QTL described here suggest a large effect of the Z chromosome on sterility (Naisbit *et al.*, 2002). However, there is no evidence that differentially expressed ovary genes are disproportionately located on the Z. This may be because ovary expressed genes more generally are under-represented on the Z chromosome, or perhaps genetic changes on the Z chromosome regulate autosomal ovary genes in *trans*, leading to a lack of overall enrichment of Z-linked differential expression. The latter hypothesis could be addressed using analysis of allele-specific expression. In summary, patterns of segregation of sterility indicates support for a large Z-effect but this is not evident in the expression data.

Haldane's Rule and the large X-effect have several potential causes. First, dominance theory predicts that the heterogametic sex is inviable or sterile primarily because of the recessive nature of incompatibility factors, which are therefore expressed on the sex chromosome in the heterogametic sex. Second, faster-X evolution predicts that incompatibilities accumulate more rapidly on the sex chromosome due to more efficient selection of recessive mutations, due to hemizygosity. Finally, the faster-male hypothesis predicts that hybrid male sterility will accumulate because male reproductive traits are subject to more rapid evolution perhaps due to sexual selection. Female-heterogametic taxa offer an opportunity to disentangle these effects to some extent because it is females rather than males that are sterile. The existence of a strong Haldane's Rule effect in female heterogametic taxa might seem to support the first two hypotheses because it is females rather than males that are sterile or inviable. I have shown, however, that there is little evidence for faster-Z evolution in these taxa (Chapter 2, "Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*"), which does not therefore provide strong support for the second hypothesis. Overall, most of the evidence in Heliconius points towards dominance theory as a primary cause for the Haldane's Rule effects that are seen in this genus,

which is also supported by theoretical predictions that BDMIs are likely to specifically involve sex chromosomes (Turelli and Moyle, 2006).

Differential expressed genes between *H. melpomene*, *H. cydno* and hybrid ovary tissue are distributed throughout the genome. The genomic landscape of differentially expressed genes is heterogeneous and this may have implications in the genetic mechanisms involved in the speciation process. Differentially expressed genes are enriched towards the ends of chromosomes. This may be simply a reflection of the fact that there is an enrichment of genic sequences towards to the periphery of chromosomes in *Heliconius* (Simon Martin, pers. comm.). Permutation tests on the total genic distribution along the reference genome need to be performed to test the significance of the reported enrichment of differentially expressed genes towards the periphery of the chromosomes. It is interesting to note, however, that recombination rates are also higher towards the ends of chromosomes in *Heliconius* (Simon Martin, pers. comm). Higher recombination rates contribute to higher efficacy of natural selection and so, potentially positively selected loci at such locations, may evolve at faster rates (Gante *et al.*, 2016).

Hybrid sterility commonly involves different genomic loci for males and females and a number of genes is likely to contribute for reproductive isolation (Orr, 1993). The different sex determination mechanisms form a continuum that correlates with the level of post-zygotic isolation observed between two species: increasing sex chromosome differentiation increases the severity of post-zygotic isolation (Lima, 2014). Through the analysis of gene expression differences between fertile and sterile butterflies I have identified genes that might be responsible for the sterility phenotype in F1 *H. cydno* x *H. melpomene* samples. I have cross-referenced such genes to a mapped sterility QTL and to genomic loci of no gene flow. Genes that are differentially expressed between the *H. cydno, H. melpomene* and the hybrids are expected to map to regions of no gene flow if they are causally involved in reproductive isolation between the two species. Future studies of this phenotype should therefore focus on those differentially expressed loci that I

have identified in regions of low gene flow as the most plausible candidate reproductive isolation genes.

Additionally, to investigate gene expression differences between *H. cydno*, *H. melpomene* and the hybrids, and its correlation to sterility in the latter, the gene expression data I generated will be analysed further. Specifically, I will determine inheritance classifications for the differentially expressed genes in the hybrids to explicitly quantify changes in expression. Moreover, I will measure the contribution of *cis* and *trans* effects to gene expression divergence in the hybrids to determine whether there is conserved regulation between the two species. A similar and balanced allelic expression between *H. cydno* and *H. melpomene* and hybrids would be an indication of conserved regulation between the two species. However, a conserved unbalanced allelic expression between *H. cydno* and *H. melpomene* and the hybrids would be the signature of parental *cis*-regulatory differences. On the other hand, a balanced allelic expression only in the hybrids would reveal parental *trans*-regulatory divergences. Establishing whether *cis*-regulatory changes are more prevalent than *trans*-regulatory changes may allow to quantify how important adaptive evolution is in the evolution of the sterility phenotype.

By quantifying differences in gene expression between fertile and sterile female reproductive tissue, and cross-referencing it with regions of restricted gene flow previously identified between *H. cydno* and *H. melpomene* crosses, I mapped loci putatively responsible for reproductive isolation between the two species. By cross-referencing two different WGS studies with *Heliconius* gene expression data I hoped to move away from the purely descriptive nature of gene expression differences that so often constitute RNA-seq studies (Roux pers. comm.; Merrill pers. comm.). A better understanding of the mechanisms shaping hybrid phenotypes may help to further elucidate the role of hybridization on the speciation process and gene mis-regulation has been shown to be a common source of incompatibilities and hybrid unfitness (Ortiz-Barrientos *et al.*, 2007; Long *et al.*, 2008; Anderson *et al.*, 2010).

Supplementary Tables

Supplementary Table S1. Sample information and sequencing statistics

Sample	Species	Sex	Tissue	Treatment	Library
AP23_Ov	H. melp	Fem	Ovary	Young	RRB02955
AP28_Ov	H. melp	Fem	Ovary	Young	RRBL00014
AP63_Ov	H. melp	Fem	Ovary	Young	RRB02956
AP67_Ov	H. melp	Fem	Ovary	Young	RRBL00005
AP21_Ov	H. melp	Fem	Ovary	Young	RRB02958
AP20_Ov	H. melp	Fem	Ovary	Young	RRB02959
AP19_Ov	H. melp	Fem	Ovary	Young	RRB02957
AP55_Ov	H. melp	Fem	Ovary	Old	RRB03012
AP77_Ov	H. melp	Fem	Ovary	Old	RRB03013
AP80_Ov	H. melp	Fem	Ovary	Old	RRB03014
AP89_Ov	H. melp	Fem	Ovary	Old	RRB03015
AP141_Ov	H. melp	Fem	Ovary	Old	RRB03016
AP142_Ov	H. melp	Fem	Ovary	Old	RRB03017
AP35_Ov	H. cydno	Fem	Ovary	Young	RRBL00006
AP94_Ov	H. cydno	Fem	Ovary	Young	RRB02962
AP34_Ov	H. cydno	Fem	Ovary	Young	RRB02963
AP37_Ov	H. cydno	Fem	Ovary	Young	RRB02960
AP71_Ov	H. cydno	Fem	Ovary	Young	RRBL00007
AP88_Ov	H. cydno	Fem	Ovary	Young	RRBL00008
AP93_Ov	H. cydno	Fem	Ovary	Young	RRB02961
AP54_Ov	H. cydno x H. melp	Fem	Ovary	Young	RRBL00009
AP38_Ov	H. cydno x H. melp	Fem	Ovary	Young	RRB03018
AP39_Ov	H. cydno x H. melp	Fem	Ovary	Young	RRB03019
AP53_Ov	H. cydno x H. melp	Fem	Ovary	Young	RRBL00015
AP58_Ov	H. cydno x H. melp	Fem	Ovary	Young	RRBL00010
AP66_Ov	H. cydno x H. melp	Fem	Ovary	Young	RRBL00011
AP65_Ov	H. cydno x H. melp	Fem	Ovary	Young	RRB03021
AP41_Ov	H. cydno x H. melp	Fem	Ovary	Young	RRB03020
AP52_Ov	H. cydno x H. melp	Fem	Ovary	Young	RRB03023
AP70_Ov	H. cydno x H. melp	Fem	Ovary	Young	RRB03022

Supplementary Table S1. Sample information and sequencing statistics (cont.)

Sample	RawReads	RawBase(G)	ErrorRate(%)
AP23_Ov	26837823	25982919	7.79
AP28_Ov	39138123	38706191	11.61
AP63_Ov	33777712	32704267	9.81
AP67_Ov	37359340	36752670	11.03
AP21_Ov	32244551	31192828	9.36
AP20_Ov	33632548	32592198	9.78
AP19_Ov	31459005	30474372	9.14
AP55_Ov	37934077	37335336	11.2
AP77_Ov	34322656	33261770	9.98
AP80_Ov	36157750	35149502	10.54
AP89_Ov	34318423	33383486	10.02
AP141_Ov	33844256	32934318	9.88
AP142_Ov	35328097	34348031	10.3
AP35_Ov	35018481	34645187	10.39
AP94_Ov	39903075	39134431	11.74
AP34_Ov	27811550	27296213	8.19
AP37_Ov	33410348	32682152	9.8
AP71_Ov	34856038	34487192	10.35
AP88_Ov	38486006	38121619	11.44
AP93_Ov	31497198	30839548	9.25
AP54_Ov	36589398	36249200	10.88
AP38_Ov	35305269	34283164	10.29
AP39_Ov	39788204	39104796	11.73
AP53_Ov	34638287	34288204	10.29
AP58_Ov	44218820	43693866	13.11
AP66_Ov	40503519	39816955	11.95
AP65_Ov	40575350	39711736	11.91
AP41_Ov	36564414	35536413	10.66
AP52_Ov	26051783	24931600	7.48
AP70_Ov	24514381	23726336	7.12

Supplementary Table S1. Sample information and sequencing statistics (cont.)

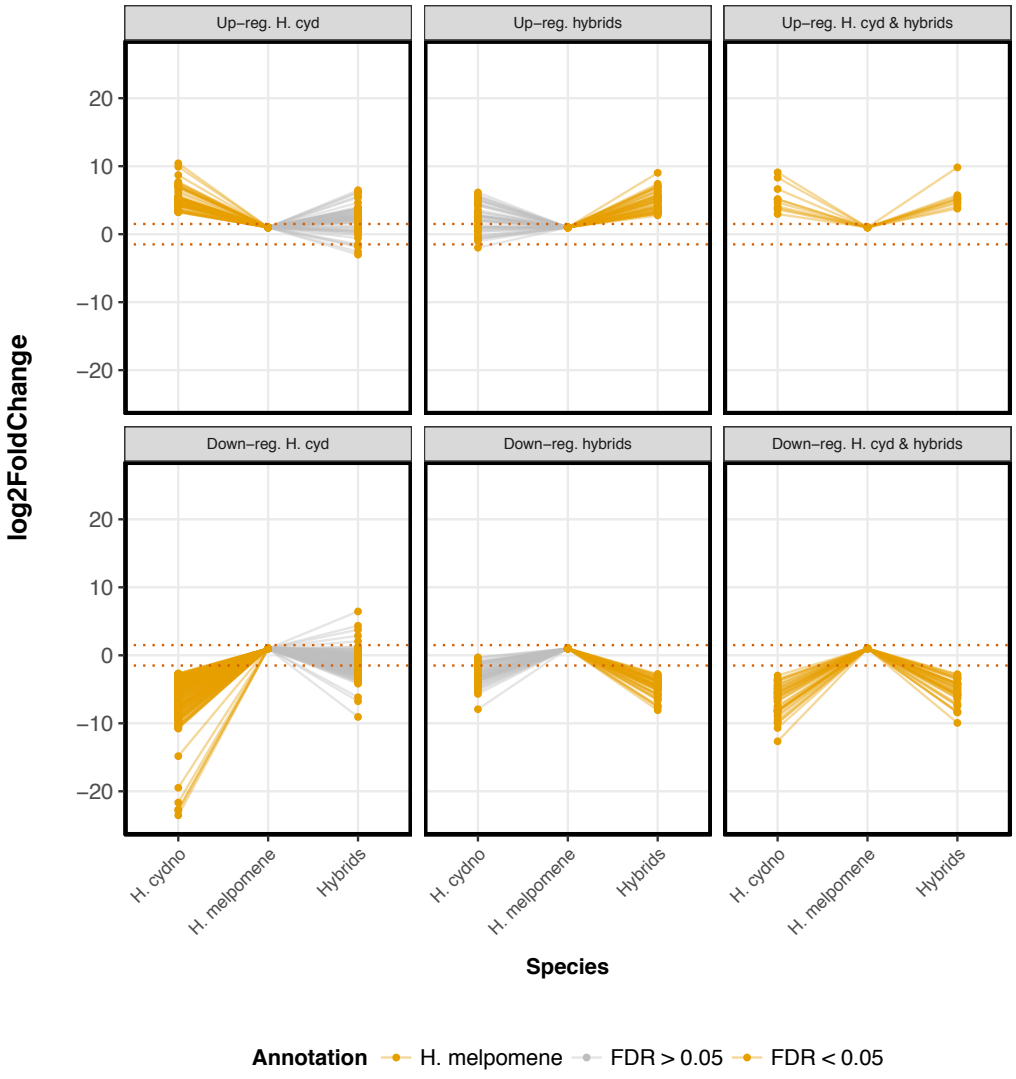
Sample	Q20 (%)	Q30 (%)	GCcont (%)	MappedReads (%)	Prop.Paired (%)
AP23_Ov	96.81	0.02	97.38	93.58	41.01
AP28_Ov	98.9	0.01	98.61	96.47	44.52
AP63_Ov	96.82	0.01	97.75	94.25	40.83
AP67_Ov	98.38	0.01	98.47	96.28	41.14
AP21_Ov	96.74	0.01	97.82	94.45	41.56
AP20_Ov	96.91	0.02	97.6	93.97	40.61
AP19_Ov	96.87	0.02	97.51	93.82	40.34
AP55_Ov	98.42	0.01	98.17	95.51	39.76
AP77_Ov	96.91	0.01	98.28	95.6	40.46
AP80_Ov	97.21	0.01	97.97	95.06	40.28
AP89_Ov	97.28	0.01	98.04	95.16	41.51
AP141_Ov	97.31	0.01	97.88	94.83	39.58
AP142_Ov	97.23	0.01	98	95.09	39.69
AP35_Ov	98.93	0.01	98.53	96.41	41.08
AP94_Ov	98.07	0.01	98.16	95.19	41.51
AP34_Ov	98.15	0.01	98.2	95.27	41.42
AP37_Ov	97.82	0.01	98.49	95.88	42.15
AP71_Ov	98.94	0.01	98.42	96.19	41.36
AP88_Ov	99.05	0.01	98.53	96.37	42.71
AP93_Ov	97.91	0.01	98.38	95.69	41.24
AP54_Ov	99.07	0.01	98.42	96.19	42.61
AP38_Ov	97.1	0.01	98.19	95.47	40.41
AP39_Ov	98.28	0.01	98.42	96.02	41.2
AP53_Ov	98.99	0.01	98.33	95.9	43.8
AP58_Ov	98.81	0.01	98.4	96.13	41.98
AP66_Ov	98.3	0.01	98.5	96.33	41.23
AP65_Ov	97.87	0.01	98.47	96.14	42.93
AP41_Ov	97.19	0.01	98.37	95.74	40.5
AP52_Ov	95.7	0.01	98.51	96.28	43.27
AP70_Ov	96.79	0.01	98.81	96.87	41.2

Supplementary Table S1. Sample information and sequencing statistics

H. cydno, *H. melpomene* and *H. cydno* x *H. melpomene* mRNA sequencing statistics. Sample ID, species, tissue, stage of collection for mRNA 150bp PE directionally sequenced reads.

Supplementary Figures

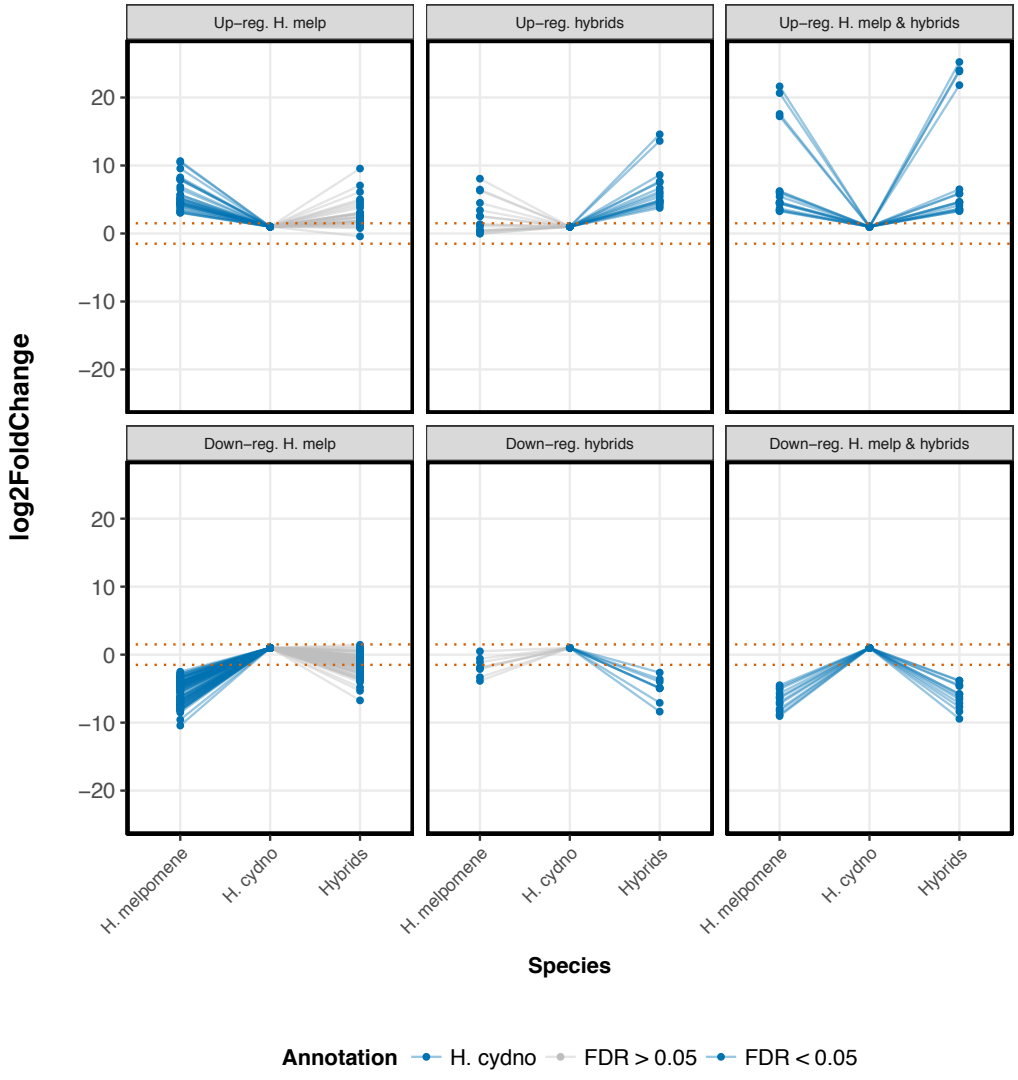
Supplementary Figure S1. Patterns of expression for the differential expressed genes identified with *H. melpomene* reference genome and annotation



Supplementary Figure S1. Patterns of expression for the differential expressed genes identified with *H. melpomene* reference genome and annotation

Genes differentially expressed identified with the analysis performed on the *H. melpomene* reference genome and annotation. Points represent the a gene in each of the three different sample groups. Gene expression for each gene calculated with *H. melpomene* samples as the baseline. Negative log2 fold values – gene expression is lower than in *H. melpomene*. Positive log2 fold values – gene expression is greater than in *H. melpomene*. Differentially expressed genes are separated by their expression patterns: 1) Up-regulated in *H. cydno*, 2) up-regulated in the hybrids, 3) up-regulated in *H. cydno* and in the hybrids, 4) down-regulated in *H. cydno*, 5) down-regulated in the hybrids, 6) down regulated in *H. cydno* and in the hybrids. Gene expression values of the three different groups (*H. cydno*, *H. melpomene* and hybrids) is linked by yellow or grey lines. Yellow lines represent significant results (FDR < 0.05 and log2 fold change > |1.5|), grey lines represent non-significant results (FDR > 0.05 and/or log2 fold change < |1.5|). Dotted red lines delineated the |1.5| log2 fold change significance threshold.

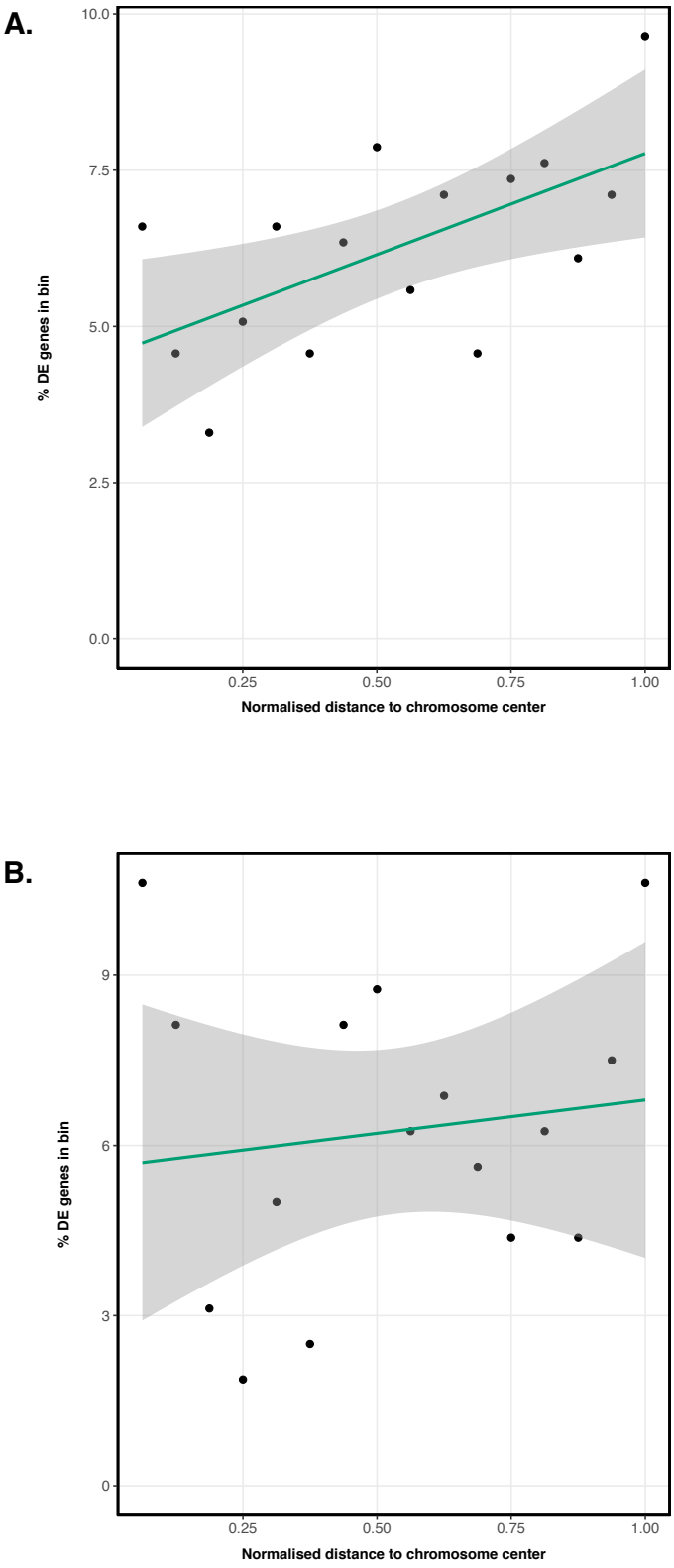
Supplementary Figure S2. Patterns of expression for the differential expressed genes identified with *H. cydno* reference genome and annotation



Supplementary Figure S2. Patterns of expression for the differential expressed genes identified with *H. cydno* reference genome and annotation

Genes differentially expressed identified with the analysis performed on the *H. cydno* reference genome and annotation. Points represent the a gene in each of the three different sample groups. Gene expression for each gene calculated with *H. cydno* samples as the baseline. Negative log2 fold values – gene expression is lower than in *H. cydno*. Positive log2 fold values – gene expression is greater than in *H. cydno*. Differentially expressed genes are separated by their expression patterns: 1) Up-regulated in *H. melpomene*, 2) up-regulated in the hybrids, 3) up-regulated in *H. melpomene* and in the hybrids, 4) down-regulated in *H. melpomene*, 5) down-regulated in the hybrids, 6) down regulated in *H. melpomene* and in the hybrids. Gene expression values of the three different groups (*H. cydno*, *H. melpomene* and hybrids) is linked by blue or grey lines. Blue lines represent significant results (FDR < 0.05 and log2 fold change > |1.5|), grey lines represent non-significant results (FDR > 0.05 and/or log2 fold change < |1.5|). Dotted red lines delineated the |1.5| log2 fold change significance threshold.

Supplementary Figure S3. Distribution of differentially expressed genes classified using the *H. cydno* and *H. melpomene* annotations separately



Supplementary Figure S3. Distribution of differentially expressed genes classified using the *H. cydno* and *H. melpomene* annotations separately

Differentially expressed genes were grouped into 16 bins along their normalised distance from the centre of the chromosome. In the x-axis, 0 is the normalised chromosome centre, 1 is the normalised chromosome end. **A.** Differentially expressed when the analysis is done with *H. cydno* reference genome/annotation; **B.** Differentially expressed when the analysis is done with the *H. melpomene* reference genome/annotation. Linear regression line fit to the correlation between total number of differentially expressed genes per bin and distance to chromosome centre.

CHAPTER 4

piRNA mediated epigenetic silencing does not underlie post-zygotic isolation between *Heliconius cydno* and *Heliconius melpomene*

Abstract

One of the major approaches to understanding speciation is through the genetic dissection of the process of reproductive isolation. Inter-specific hybridization can lead to genomic stress in the form of chromosomal rearrangements, changes in recombination and mutation rates, changes in gene expression and DNA methylation, or activation of transposable elements (TEs). To assess the role of TEs in inter-specific F1 *Heliconius cydno* x *Heliconius melpomene* female sterility, I performed a TE transcriptomic analysis in the ovaries of sterile F1 females and the fertile *H. cydno* and *H. melpomene*. I found 14 TEs mis-regulated between F1 female hybrids and *H. melpomene*. *Piwi*-interacting RNAs (piRNAs) are responsible for the silencing of TEs and, in my crosses, there are no piRNA gene expression differences between the sterile F1s and the parental species. The piRNA pools between the three populations are equivalent and F1 hybrids produce piRNAs to silence the mis-regulated TEs. I conclude that neither functional divergence of the piRNA pathway, deregulation of specific TE families nor the absence of specific piRNAs is likely to explain the sterility phenotype observed in F1 females but that TE expression may impact neighbouring protein coding gene expression.

Introduction

All classes and types of eukaryotic TEs have been identified in insects and the repetitive nature of TEs can be used for their discovery (Lavoie *et al.*, 2013). There is a large body of literature highlighting the role of selfish genetic element divergence between populations in hybrid incompatibilities and speciation (Kidwell *et al.*, 1977; Ortiz-Barrientos *et al.*, 2007; Presgraves, 2010; Kelleher *et al.*, 2012; Hill *et al.*, 2016). Selfish genetic elements, such as TEs, can spread and lead to rapid evolution of genetic differences between closely related populations. When two populations with different genetic backgrounds hybridize, TE de-repression is common. Since TE proliferation is deleterious, mechanisms to control TE mobilization in the germline, halting the spread of novel TEs to the progeny, are crucial for the stability of the genome (Czech and Hannon, 2016).

Most animal genomes have an active piRNA pathway for TE silencing. piRNAs, a type of small RNA, are a 23 to 31 nucleotides long. The piRNA pathway is a conserved maternally inherited defence mechanism (genomic immune system) that acts against the deleterious effects of transposons. piRNAs homologous to TEs are sequestered in the egg's cytoplasm and target TEs for mRNA degradation. piRNA templates can be found within discrete genomic clusters. Transcription through these clusters produces single stranded piRNA precursors that are cleaved to produce primary piRNAs. In some species, the Ping-Pong cycle is required for primary piRNAs to recognise their complementary targets and for the recruitment PIWI proteins. In such cases, primary piRNAs go through the Ping-Pong amplification cycle becoming secondary piRNAs. Alternatively, piRNAs can also be produced by cleavage of piRNA cluster transcripts processed during secondary piRNA biogenesis. Regardless, if piRNAs are present, TE activity is reduced and DNA damage halted (Czech and Hannon, 2016; Khanduja *et al.*, 2016).

Studies describing enhanced TE expression suggest that TE over-expression in hybrids is caused by TE silencing breakdown (Kelleher *et al.*, 2012; Lopez-Maestre *et al.*, 2017). There are two different explanations for the observed breakdown and TE de-repression could be due to: 1) maternal cytotype failure; 2) global failure of the piRNA pathway (Romero-Soriano et al 2017).

For the maternal cytotype failure hypothesis, females lacking piRNAs that target a specific TE are mated to males that contain an active copy of that TE. This occurs due to differential TE insertion into piRNA clusters in the different lineages, without any change in protein coding genes (Grentzinger *et al.*, 2012). Since maternally deposited piRNAs are necessary to initiate TE silencing in the progeny, these foreign transposons are not silenced and can dramatically reduce the fitness of the new host individual. In such cases, the progeny has increased mutation and recombination rates, sterility and dysgenic (small) gonads due to DNA damage caused by de-repression and active transposition of TEs in the germline (Kidwell *et al.*, 1977; Kidwell, 1983; Hill *et al.*, 2016).

Alternatively, TE de-repression in hybrids due to a global failure of the piRNA pathway can result from adaptive divergence of piRNA pathway genes (Obbard, Welch, *et al.*, 2009; Kelleher *et al.*, 2012). TEs are obvious candidates to drive adaptive evolution of piRNA-effector proteins. On one hand, antagonist evolution between TEs and the piRNA pathway might be analogous to what has been observed between viruses and the small interfering RNA pathway, where host proteins must adapt to avoid functional disruption by viral proteins. On another hand, piRNA proteins could evolve quickly to respond to changes in the content of the TE pool of the host genome (Singh *et al.*, 2009). In a global failure of the piRNA pathway scenario, protein divergence plays a role in the evolution of host genome defence against TEs and, consequently, post-zygotic isolation (Begun *et al.*, 2007; Obbard, Gordon, *et al.*, 2009).

These two alternative hypotheses are not mutually exclusive and can occur simultaneously in certain *Drosophila* crosses (Romero-Soriano et al 2017). However, generally, the maternal cytotype failure scenario is more common between intra-specific crosses of *Drosophila*; and the global failure of the piRNA pathway scenario between inter-specific crosses (Bucheton *et al.*, 1976; Blackman *et al.*, 1987; Hill *et al.*, 2016; Czech and Hannon, 2016).

Until recently, our knowledge of piRNA mediated silencing was largely restricted to *Drosophila*. With next-generation sequencing it is now possible to investigate sRNA pathway in organisms where RNAi silencing is not possible (Lewis *et al.* 2017). In Drosophila, piRNAs are restricted to the germline, but a recent study shows that in *Heliconius* piRNAs are present in both the soma and germline of males and females (Lewis *et al.* 2017). This suggests that, in *Heliconius*, piRNAs target not only germline TEs but also somatic TEs, viruses and mRNAs; and confirms how studies of one organism do not always accurately represent biological diversity and we should be cautious when making generalisations.

Using the *H. melpomene* v1.0 has been estimated that TEs comprise roughly 25% of the *H. melpomene* genome (Lavoie *et al.*, 2013), the piRNA effector genes have been identified, and piRNA read-length distributions are described (Lewis *et al.* 2017). *Heliconius* piRNAs display the features of piRNA biogenesis and amplification having a 5' uracil bias, 5' nucleotide complementarity between piRNAs from opposite strands (i.e. Ping-Pong signature), and resistance to oxidation by sodium periodate (i.e. with a 2′-*O*-methyl modification at their 3′ ends) (Lewis *et al.* 2017).

Here, I investigate whether epigenetic silencing mechanisms could underlie post-zygotic isolation between *H. cydno* and *H. melpomene*, two hybridizing sympatric neotropical butterfly species that differ in their ecology, mimicry patterns and mate preferences (Jiggins *et al.*, 2001; Merrill *et al.*, 2012; 2013). Additionally to pre-zygotic barriers to gene flow, there are also post-zygotic barriers including increased predation and F1 sterility (Naisbit *et al.*, 2002;

Merrill *et al.*, 2012). Hybrid F1 female progeny of the *H. cydno* x *H. melpomene* cross is always sterile (Naisbit *et al.*, 2002) but able to develop some ovarian tissue and, occasionally, oocytes (Chapter 3 for detailed Results). I quantify TE expression in *H. cydno* x *H. melpomene* female hybrids, examine whether the piRNA silencing pathway is functional and describe the piRNA pool. By measuring TE expression and piRNAs in the hybrids, *H. cydno* and *H. melpomene*, I test whether TE silencing breakdown is correlated to F1 hybrid sterility.

Materials and Methods

Intra- and inter-specific crosses of *H. cydno* and *H. melpomene*

Crosses were carried between *H. cydno chioneus* and *H. melpomene rosina* at the Smithsonian Tropical Research Institute insectaries in Gamboa, Panama (9º08'N 7º42'W). All mothers of broods are virgin insectary-bred females. Fathers of broods are wild caught individuals collected along Pipeline Road in the Soberanía National Park (9º87'N 7º96'W). Intra- and inter-specific crosses were carried out as described in Chapter 3, "Sterility in *Heliconius cydno* x *Heliconius melpomene* F1 female hybrids: a phenotypic and gene expression study of hybrid incompatibilities".

I collected eggs every day for both inter-specific and intra-specific crosses and laying females had access to *Psiguria* flowers; *Lantana camara*; and artificial feeders as described in Chapter 3, "Sterility in *Heliconius cydno* x *Heliconius melpomene* F1 female hybrids: a phenotypic and gene expression study of hybrid incompatibilities". I kept the collected eggs and catterpillars were treated also as described in Chapter 3. When a female emerged I either: 1) took it back to the insectaries and allowed to mature (Chapter 3 for details on phenotypic study); or dissected it for coding (Supplementary Table S1) and non-coding (Supplementary Table S2) transcript analysis of ovary tissue.

H. cydno, H. melpomene and F1 female hybrid tissue collection for coding and non-coding transcript abundance

I dissected ovary tissue 1h to 3h after eclosion for *H. cydno, H. melpomene* and F1 hybrid females that pupated and emerged in the laboratory (Appendix B, Protocol for dissections of the reproductive tract for total RNA extraction). The ovary tissue from newly eclosed females was sequenced to quantify transcript abundance of coding and non-coding elements (mRNA and sRNA sequencing sRNA refers to all species of sRNA (piRNA, miRNA and siRNA). sRNA will be used throughout the chapter when discussing sequencing and the overall sRNA landscape. piRNA species, the focus of this study, will be used when specifically discussing results related to this species of sRNA. I also dissected 20-day old *H. melpomene* ovary tissue to quantify transcript abundance of coding elements (mRNA sequencing) (Chapter 3 for detailed Methods & Results).

Total RNA extraction for mRNA sequencing

For seven newly eclosed *H. melpomene* ovaries, seven newly eclosed *H. cydno* ovaries, ten newly eclosed hybrid ovaries, and six 20-day old *H. melpomene* ovaries total RNA was extracted with a combined guanidium thiocyanate-phenol-chloroform and silica matrix protocol using TRIzol (Invitrogen, Carlsbad, CA), RNeasy columns (Qiagen, Valencia, CA) and DNaseI (Ambion, Naugatuck, CT) (Appendix C, Total RNA extraction protocol for mRNA sequencing) (Chapter 3 for detailed Methods & Results). mRNA isolated from total RNA via poly-A pull-down, directional cDNA libraries and 150bp paired-end sequencing done by Novogene Bioinfomatics Technologies (~30M reads/sample) (Hong Kong, China) (Appendix C, Total RNA extraction protocol for mRNA sequencing; Supplementary Table S1).

Total RNA extraction for sRNA sequencing

For 7 newly eclosed *H. melpomene*, 4 newly eclosed *H. cydno*, and 5 newly eclosed hybrid ovaries, total RNA was extracted with an isopropanol-chloroform extraction after homogenizing the tissue in TRIzol (Invitrogen, Carlsbad, CA) (Ashe et al., 2013). RNA integrity was checked using the NanoDrop Nucleic Acid Quantification (ThermoFisher, Waltham, MA, USA) (Appendix D, Total RNA extraction protocol for sRNA sequencing). sRNAs library preparation, cDNA sequencing after adaptor ligation, reverse transcription, PCR enrichment, purification and size selection was done in Novogene Bioinfomatics Technologies (50bp single-end reads, ~40M reads/sample) (Hong Kong, China) (Supplementary Table S2). To sequence all sRNAs in a 5'-independent manner, we removed 5' triphosphates with 5' polyphosphatase (Epicentre/Illumina, Madison, WI, USA).

H. cydno and *H. melpomene* reference genome and annotation

The *H. cydno* reference genome used throughout this study was generated from a haplotypic trio assembly using progressiveCactus and Ragout (Chapter 3 for detailed Methods & Results) (Paten, Diekhans, *et al.*, 2011; Paten, Earl, *et al.*, 2011; Davey *et al.*, 2016). The *H. melpomene* reference genome used throughout this study is the published version of Hmel2 (Davey *et al.*, 2016). The *H. melpomene* annotation used here is an upgrade of the released Hmel2 annotation. This new *H. melpomene* annotation is publicly available at LepBase (Challis *et al.* BioRxiv preprint). The *H. melpomene* annotation was transferred to the *H. cydno* genome assembly using RATT and the latter was used to count *H. cydno* transcripts (Chapter 2 & 3 for detailed Methods & Results, *H. cydno* reference genome and annotation and *H. melpomene* reference annotation files accessible from https://www.dropbox.com/sh/5krc7kn3u0oviwj/AADHTlQsoxQCnqZnivatNdRba?dl=0).

Transposable element annotation

RepeatMasker (v4.0.6) was used with the *Metazoa* library to identify homologs to any previously identified metazoan TEs in the *H. cydno* reference genome (Smith et al. 2013). In addition, RepeatModeler (v1.0.8) was used to produce *de novo* Hidden Markov Model for TEs in each genome (Smit et al. 2008). RepeatMasker was posteriorly run using this HMM to identify TEs without sufficient homology to previously identified metazoan TEs. The two *H. cydno* TE annotations were combined to generate a single TE annotation file. All TE annotations smaller than 100 nucleotides long were discarded from the *H. cydno* TE annotation. For *H. melpomene*, the *H. melpomene* TE annotation generated by Lewis et al. (2017) was used. This *H. melpomene* TE annotation was generated as it is described above for *H. cydno*. The source code for this analysis is accessible in GitHub (https://github.com/SamuelHLewis/TEAnnotator).

TE transcript count and differential abundance

I trimmed mRNA-seq reads with default settings using Trim Galore to 1) remove adapter sequences; and 2) low quality reads from the RNA-seq mate pairs (N>10%; Qphred<5 in over 50% reads) (https://github.com/FelixKrueger/TrimGalore). Adapter sequences: 5' adapter: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT; 3' adapter: 5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG. To analyse TE expression on the samples (Table 1) I used TEtools (v1.0.0) which takes into account TE sequence diversity of the reference genome (Lerat *et al.*, 2017). Using the TEcount module of TEtools (-RNApair & default vales), I mapped mRNA-seq data from *H. cydno, H. melpomene* and F1 female hybrid ovary tissue to the 1) *H. cydno* TE annotation library; and to the 2) *H. melpomene* TE annotation library. Bowtie2

(default values, v2.2.4) is a dependence of TEtools for mapping (Langmead and Salzberg, 2012). The pipeline has been tested in several studies since its publication and it has been consistently shown to perform well (Jakšić *et al.*, 2017; Romero-Soriano *et al.*, 2017; Ryazansky *et al.* 2017). Estimation of variance-mean dependence from the count data was performed with the DESeq2 (v1.14.1) of Bioconductor (v3.4) in the R software environment (v3.2.5) to calculate TE differential abundance using the constructor function DESeqDataSetFromHTSeqCount(design=~batch+species). I built the result tables using the DESeq2 results() function (options: betaPrior=false, test=Wald) (Love *et al.*, 2014). I filtered the results as in Walters *et al.* (2015) with log2 fold significance threshold > |1.5| and FDR < 0.05 (options: lcfThreshold=1.5, altHypothesis="greaterAbs", alpha=0.05) (Walters *et al.*, 2015).

Samples	Ovary	
	3h after eclosion	20 days after eclosion
 ~34M reads/sample ♀	7 x 150 PE	6 x 150 PE
 ~34M reads/sample ♀	7 x 150 PE	NA
 ~34M reads/sample ♀	10 x 150 PE	NA

Table 1. Samples used to calculate TE transcript abundance differences

Total number of samples used to estimate gene expression differences between *H. cydno*, *H. melpomene* and *H. cydno* X *H. melpomene*. Average read number for each group is reported as well as time of dissection: 24 samples were dissection ~3h after eclosion and 6

samples 20 days after eclosion. PE refers to paired-end RNAseq reads.

piRNA genes transcript abundance

The three *Piwi* proteins previously identified by Lewis *et al.* 2017 in *H. melpomene* have a 1-1 correspondence with *H. cydno.* To quantify piRNA gene transcript abundance I used the trimmed and quality filtered mRNA reads from the *H. cydno, H. melpomene* and hybrid females. I aligned the fastq reads to the gene sequences from *H. melpomene* and *H. cydno* annotation file with HISAT2 (Kim *et al.*, 2015). I mapped reads with HISAT2 done with default mapping parameters and calculated summary mapping statistics with samtools flagstat (v1.2) (Li *et al.*, 2009). I used htseq-count to count how many aligned sequencing reads mapped to each genic feature (HTSeq v0.6.1; python v2.7.10; option: -m union) (Anders *et al.*, 2015). Estimation of variance-mean dependence from the count data, differential expression analysis and filtering were all performed with DESeq2 (Love *et al.*, 2014) (Chapter 3 for detailed Methods and Results). piRNA gene expression abundances were selected from each one of the samples considering transcriptome average.

sRNA analysis

To characterise sRNAs derived from the genome I used the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to screen out small RNA reads with >10% positions with a Qphred score <20 and cutadapt to trim adapter sequences from reads (Martin, 2012). I mapped sRNAs from the *H.* cydno, *H. melpomene* and hybrids to the reference TE libraries for *H. melpomene* and *H. cydno* using Bowtie2 (--fast mode, v2.2.4) (Langmead and Salzberg, 2012) and quantified the length distribution, base composition and strand distribution

of sRNAs using custom Python scripts (GitHub https://github.com/SamuelHLewis/sRNAplot).

First, I considered all the sRNA sequences targeting all TEs considering unique sRNA sequences only. Next, I characterised sRNAs targeting each family of TE separately: DNA TEs; LINE; SINE; RC; LTR and unclassified. Finally, I characterised sRNAs mapping to each one of the over-expressed TEs separately. To characterise sRNAs mapping to: 1) each TE family separately; and to 2) each de-repressed TE I used bedtools getfasta to extract TE sequences for the genome in a strand-specific manner (Quinlan and Hall, 2010). sRNAs were mapped as it is described above for all TEs but, this time, considering all sRNA sequences (Lewis et al. 2017).

Predicting protein class and domains for genes flanking under-expressed TEs

I used InterProScan (v5.18.57.0) (options –t n –goterms) to scan genic sequences within 2 kb of under-expressed TEs against the InterPro signatures (Wang *et al.*, 2013). InterPro signatures are predictive models provided by several different databases such as Gene3D, InterPro, Pfam, PRINTS, SUPERFAM, PROSITE and PANTHER. This allowed for functional analysis of proteins by classifying them into families and predicting domains (Mitchell *et al.*, 2015).

Results

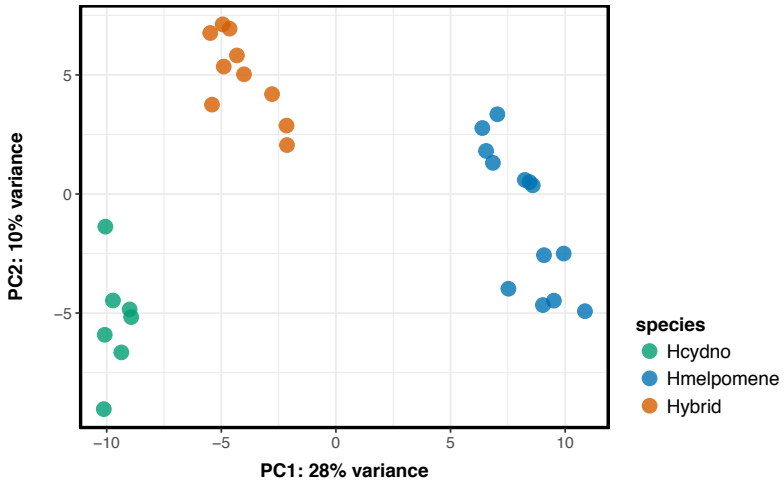
No global TE de-repression in F1 female hybrids

Firstly, I wanted to investigate whether there was global TE de-repression in the ovaries of F1 females akin to what is observed in F1 progeny from inter-specific *Drosophila* crosses. I probed such TE for de-regulation in the hybrids

by analysing ovarian mRNA from newly eclosed *H. melpomene*, *H. cydno* and hybrid females; and 20-day old *H. melpomene* females.

The samples cluster by species when TEs are mapped to the reference *H. cydno* repeat library. However, when the samples are mapped to the reference *H. melpomene* repeat library, the *H. cydno* and the hybrid samples do not form clear separate clusters. The lack of separation between the *H. cydno* and hybrid samples when the samples are mapped to the *H. melpomene* reference indicates that the TE landscape is more different between the *H. melpomene* and the hybrids than between *H. cydno* and the hybrids. However, it is important to note that, when the samples are mapped to the *H. melpomene* genome there is a clear *H. melpomene* sample outlier and this sample may be sufficient to drive the clustering of the others (Figure 1). Regardless, when each TE is analysed individually it is still apparent that the *H. cydno* and the hybrid samples are more similar in TE expression than the *H. melpomene* and the hybrid (details below, Figure 2).

A.



B.

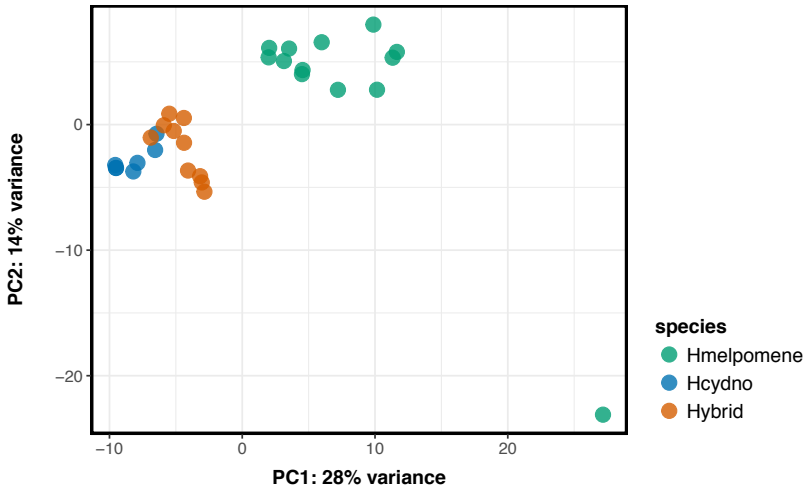


Figure 1. Principal component analysis of TE transcript abundance counted after mapping to *H. cydno* and to the *H. melpomene* reference repeat annotation libraries

A. PCA of TE transcript abundance in each sample counted against the reference *H. cydno* repeat annotation library. All the samples clearly separate by species and the hybrids cluster separately from *H.*

cydno and *H. melpomene*. **B.** PCA of TE transcript abundance in each sample counted against the reference *H. melpomene* repeat annotation library. *H. melpomene* samples clearly separate from *H. cydno* and the hybrids but there is less clear separation between *H. cydno* and the hybrids.

The *H. melpomene* genome seems to have a larger TE content than the *H. cydno* genome (Table 2) and the clustering of the *H. cydno* and hybrids samples when mapping to the *H. cydno* genome may simply reflect this (Figure 1B). On one hand, the *H. cydno* genome reference genome is a transfer from the *H. melpomene* genome reference and some repeat regions are likely to not have been transferred (Chapter 3 for detailed results). Specifically, the *H. cydno* reference genome size is smaller than the *H. melpomene* and this might reflect the fact that repeat regions did not get assembled corrected during the transfer (Table 2). On the other hand, TE content varies significantly among insect genomes. For example, in *Apis mellifera* only 1% of the genome is repeat elements (Honeybee Genome Sequencing Consortium, 2006), contrasting to 16 % in *Anopheles gambiae* (Holt *et al*., 2002) or 47% in *Bombyx mori* (Osanai-Futahashi *et al*., 2008). Between different species of *Drosophila* TE content varies from 2.7% to 25% (Drosophila 12 Genomes Consortium *et al*., 2007) and so it is plausible for *H. cydno* to harbour less repeats than *H. melpomene*. Regardless, in the future, TE annotation in *H. cydno* needs to be revisited to distinguish between the two hypotheses.

	H. melpomene	*H. cydno*
Unique TE number	58	14
Unique TE length	596 220 bp	10 728 bp
Total TE number	304	260
Total TE length	36.9 Mb	4.9 Mb
Total genome size	275 Mb	261 Mb
% TE in genome	13.41 %	1.9 %

Table 2. Summary statistics of TE annotation in *H. cydno* and *H. melpomene* genomes

Unique TE number refers to the number of unique TEs that are found in the *H. cydno* genome but not in the *H. melpomene* genome and vice-versa. *Unique TE length* is the total number of base-pairs that the *Unqiue* TEs span in either the *H. cydno* or the *H. melpomene* genome. *Total number of TE* refers to both unique and shared TEs in the *H. melpomene* and *H. cydno* genome. *Total TE length* refers to the length of TEs in the genome annotation and *% of TE in genome* the percentage of the reference genome that is composed of TEs.

The degree of correlation between TE transcript abundance and transposition remains unknown. However, transcript abundance is a direct indicator of the efficacy of transcriptional and post-transcriptional silencing and, therefore, it provides a strong indicator of piRNA-mediated silencing. Differential expression analysis of the TE-derived mRNAs mapped to the reference repeat annotation libraries for *H. cydno* and *H. melpomene* shows there is not a widespread de-repression of TEs in interspecific hybrids relative to *H. cydno* and *H. melpomene*. There are no TEs mis-expressed in the hybrids when the read pairs are counted using the *H. cydno* repeat annotation. The TEs that are significantly over/under-expressed when the read pairs are counted using the

H. cydno reference are only present in *H. melpomene* samples (Figure 2A, Supplementary Figure S1).

There are, however, 5 over-expressed and 9 under-expressed TEs in F1 female hybrids when the read pairs are counted using the reference *H. melpomene* repeat annotation. Note that the father of the inter-specific cross is an *H. melpomene* male (Figure 2B, Table 3, Figure 3, Figure 4, Supplementary Figure S2).

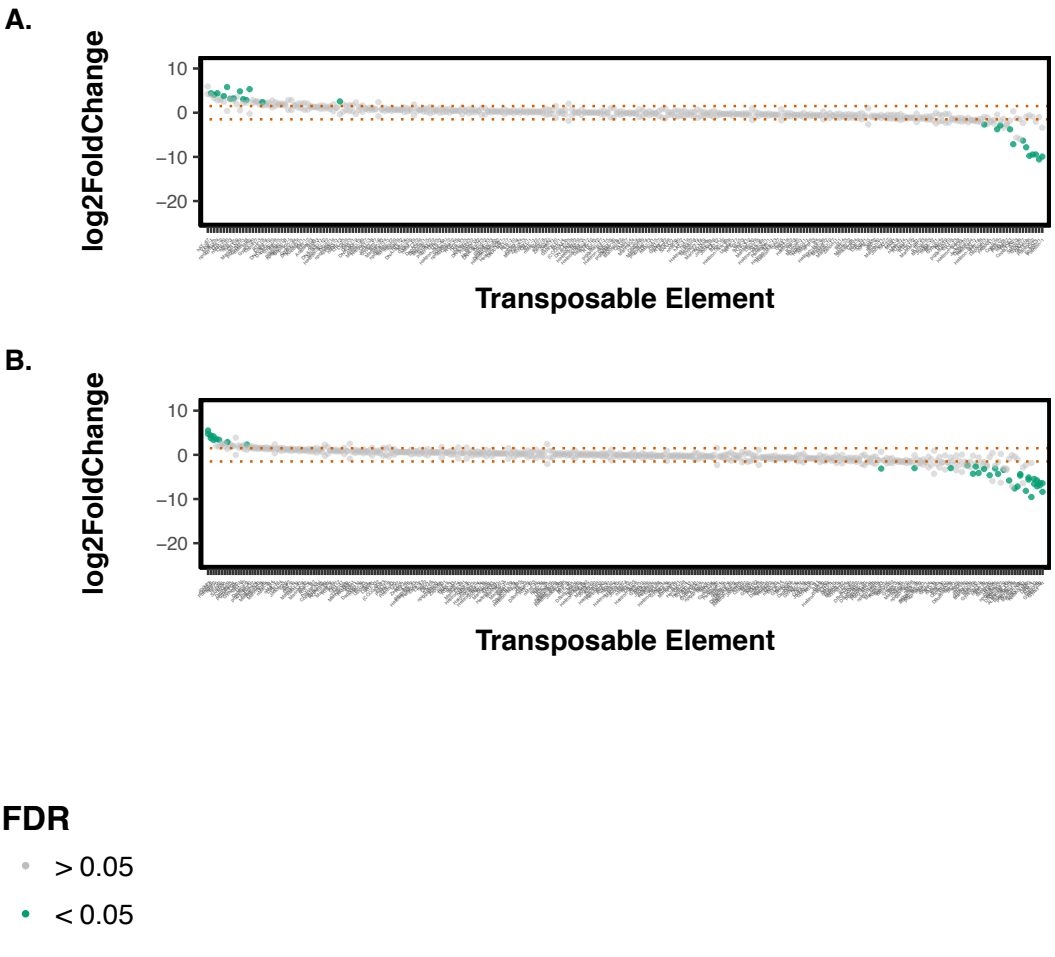


Figure 2. TEs transcript abundance differences between the hybrids and *H. cydno* or *H. melpomene*

A positive log2 fold change in A vs. B means that the number of transcripts is higher in A in comparison to B. x axis: log2 fold change; y axis: individual transposable elements ordered by log2 fold change value. Significant log2 fold changes in transcript abundance between the groups are represented in green (FDR<0.05). Dashed lines indicate l1.5l log2 fold change threshold. **A.** TE transcript abundance differences between *H. cydno*, *H. melpomene* and the hybrids using *H. cydno* TE abundances as the baseline. There are no TE transcript abundance differences in the hybrids compared to *H.* cydno and all the significantly different TEs (green dots) are from expression differences in *H. melpomene* being higher or lower than in *H. cydno*. **B.** TE transcript abundance differences between *H. melpomene*, *H. cydno* and the hybrids using *H. melpomene* TE abundances as the baseline. There are TE both over- and under- expressed in *H. cydno* and hybrids when *H. melpomene* is used as the baseline – there are significantly different TEs (green dots) from *H. cydno* and *H. melpomene* samples. There are 5 elements that are significantly more abundant in the hybrids than in *H. melpomene* and 9 that are significantly less abundant in the hybrids than in *H. melpomene* (FDR<0.05).

Transposable element name	log2 Fold change	FDR	Type	Class
BEL-2	3.41	0.004	LTR retro-transposon	I
Copia-6	3.78	0.001	LTR retro-transposon	I
Jockey	5.50	0.0003	LINE, Non-LTR retro-transposon	I

nPIF-5	3.57	0.00007	PIF/Harbinger DNA transposon	II
Polinton-2	3.36	0.04	Maverick DNA transposon	II
ALRY-MAJOR	-5.8	0.005	Major-repeat unit	II
Gypsy-10	-2.39	0.02	LTR retro-transposon	I
Gypsy-299	-6.9	0.007	LTR retro-transposon	I
Gypsy-41	-6.43	0.003	LTR retro-transposon	I
Gypsy-6	-5.78	5.53e-8	LTR retro-transposon	I
LTR-10	-5.45	6.31e-6	LTR retro-transposon	I
MAG	-5.61	8.58e-77	Gypsy LTR retro-transposon	I
Mariner-N29	-4.35	0.003	Mariner DNA transposon	II
nMar-13	-2.64	0.008	Mariner DNA transposon	II

Table 3. log2 fold changes and false discovery rates of TE with different abundances in the F1 female hybrids

Transposable elements that differ significantly in abundance in the hybrids samples.

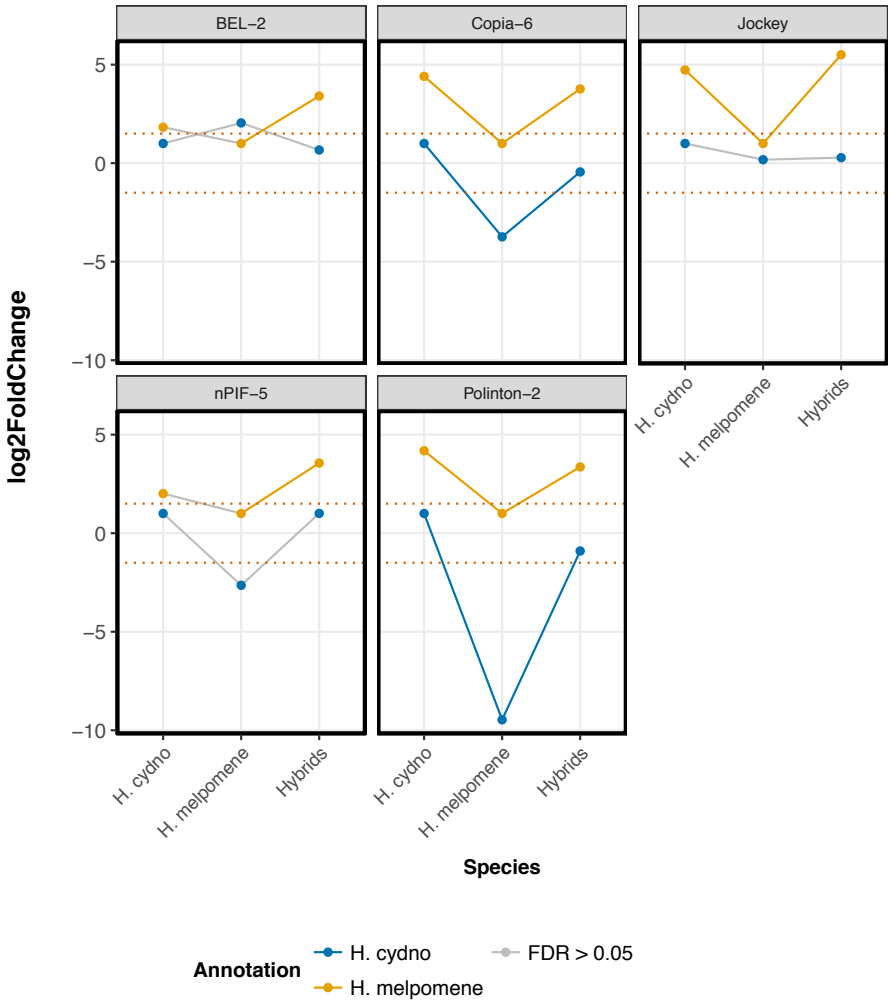


Figure 3. Pattern of TEs with higher expression in F1 female hybrids

There are 5 elements that exhibit higher expression in the hybrids. Analysis performed on the *H. cydno* (blue) and the *H. melpomene* (yellow) reference repeat annotations. Points represent a TE in each of the three different sample groups. Expression for each TE calculated with *H. cydno* and the *H. melpomene* samples as the baseline. Negative log2 fold values – gene expression is lower than in baseline. Positive log2 fold values – gene expression is greater than in baseline. TE expression values of the three different groups (*H. cydno*, *H.*

melpomene and hybrids) are linked by blue, yellow or grey lines. Blue/yellow lines represent significant results (FDR < 0.05 and log2 fold change > |1.5|), grey lines represent non-significant results (FDR > 0.05 and/or log2 fold change < |1.5|). Dotted red lines delineated the |1.5| log2 fold change significance threshold.

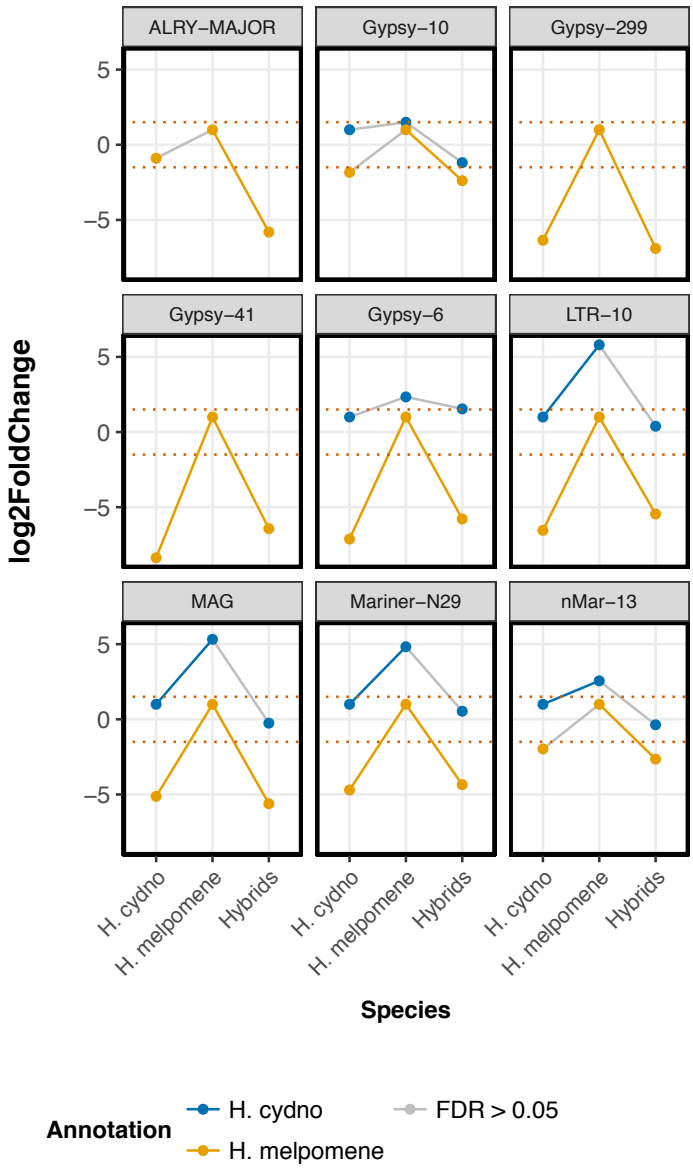


Figure 4. Pattern of TEs less abundant in F1 female hybrids

There are 9 elements that exhibit lower expression in the hybrids. Analysis performed on the *H. cydno* (blue) and the *H. melpomene* (yellow) reference repeat annotations. Points represent a TE in each of the three different sample groups. Expression for each TE calculated with *H. cydno* and the *H. melpomene* samples as the baseline. Negative log2 fold values – gene expression is lower than in baseline. Positive log2 fold values – gene expression is greater than in baseline. TE expression values of the three different groups (*H. cydno*, *H. melpomene* and hybrids) are linked by blue, yellow or grey lines. Blue/yellow lines represent significant results (FDR < 0.05 and log2 fold change > |1.5|), grey lines represent non-significant results (FDR > 0.05 and/or log2 fold change < |1.5|). Dotted red lines delineated the |1.5| log2 fold change significance threshold. ALRY-MAJOR, Gypsy-299 and -41 are not present in the *H. cydno* repeat annotation.

piRNA pathway genes are expressed at a similar level in the *H. cydno*, *H. melpomene* and hybrids

As there is no global TE de-repression in F1 female hybrids, the few over-expressed TEs observed in inter-specific hybrids are unlikely to reflect adaptive divergence of piRNA pathway genes. To confirm this I have quantified the expression of all protein coding genes and then selected the piRNA pathway genes for *H. cydno, H. melpomene* and hybrid samples to probe whether there is differential expression between the three different groups for these genes (considering the whole transcriptome landscape). None of the three piRNA pathway genes are differential expressed between *H. cydno*, *H. melpomene* and the hybrids (FDR > 0.05) but there are other protein coding genes which are (see Chapter 3 for detailed analysis on differential expressed in protein coding genes, Figure 5).

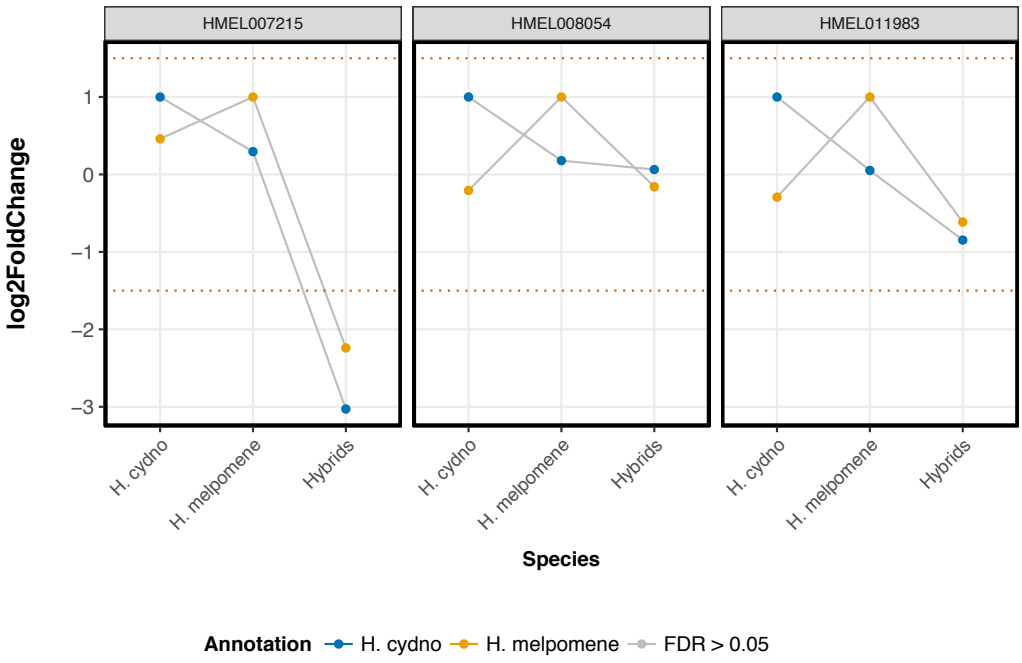


Figure 5. piRNA gene transcripts do not differ significantly between *H. cydno*, *H. melpomene* and the hybrids

Differential expression analysis on protein coding genes shows that none of the three identified piRNA genes are differentially expressed between the groups. Analysis performed on the *H. cydno* (blue) and the *H. melpomene* (yellow) reference genome and annotation. Points represent a gene in each of the three different sample groups. Gene expression for each gene calculated with *H. cydno* and the *H. melpomene* samples as the baseline. Negative log2 fold values – gene expression is lower than in baseline. Positive log2 fold values – gene expression is greater than in baseline. Gene expression values of the three different groups (*H. cydno*, *H. melpomene* and hybrids) are linked by blue, yellow or grey lines. Blue/yellow lines represent significant results (FDR < 0.05 and log2 fold change > |1.5|), grey lines represent non-significant results (FDR > 0.05 and/or log2 fold change < |1.5|).

Dotted red lines delineated the |1.5| log2 fold change significance threshold.

sRNA pools from *H. cydno*, *melpomene* and F1 inter-specific female hybrids show similar read size distributions

I sequenced the small RNA from the parents and hybrids to further establish whether the sRNA pathway is functional in F1 hybrids. sRNA pools from inter-specific hybrids have read length distributions identical to *H. melpomene* and *H. cydno* and show the characteristic Ping-Pong signature of adenine at the tenth position (Figure 6 for overview, Supplementary Figure S3, S4, S5 and S6 for all the samples). These distributions further indicate that the piRNA pathway genes in the hybrids are functional as sRNAs are present with equivalent read distribution lengths in the *H. cydno*, *H. melpomene* and the hybrids.

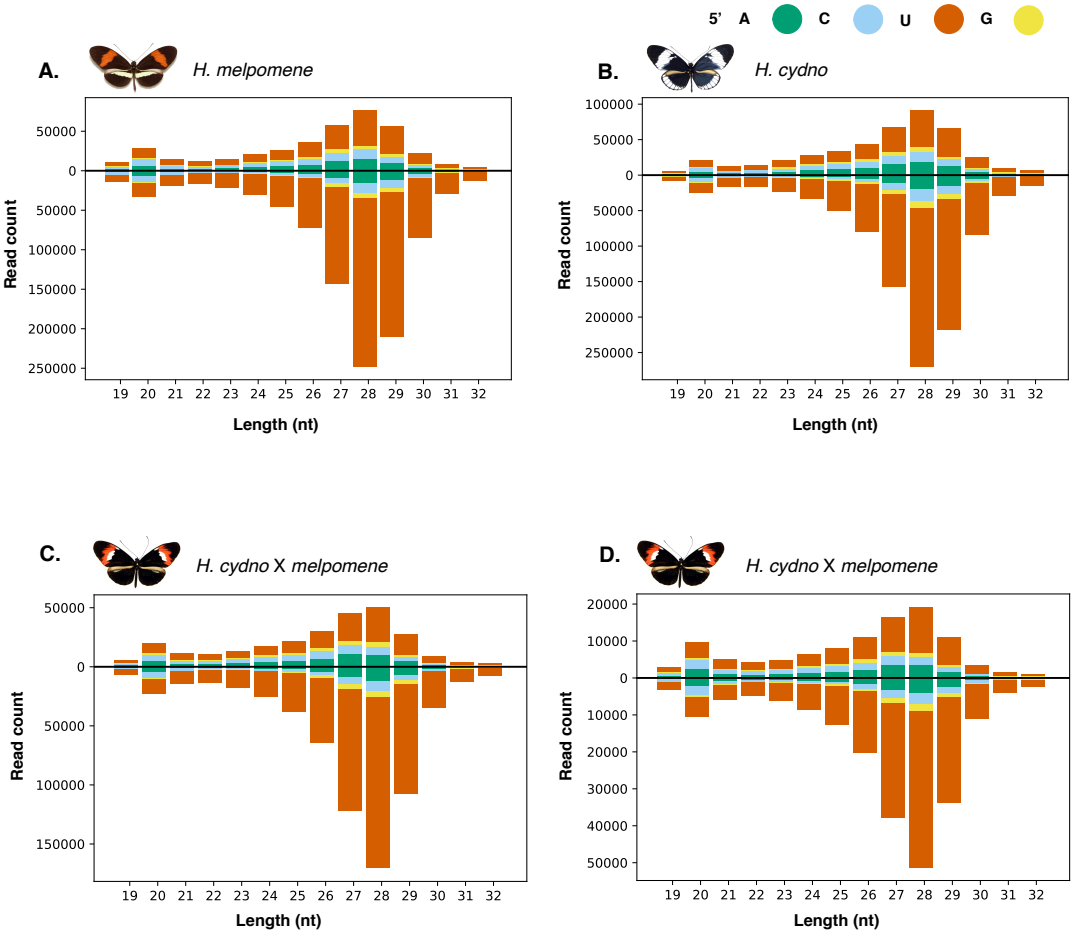


Figure 6. sRNA pool for *H. melpomene*, *H. cydno* and hybrid – overview

sRNA read distribution for **A.** *H. melpomene* (sample AP4) sRNA pool mapped to *H. melpomene* TE annotation; **B.** *H. cydno* (sample AP5) sRNA pool mapped to *H. cydno* TE annotation; **C.** *H. cydno* X *melpomene* hybrid (sample AP50) sRNA pool mapped to *H. melpomene* TE annotation; **D.** *H. cydno* X *melpomene* hybrid (sample AP50) sRNA pool mapped to *H. cydno* TE annotation. y axis: read count, x axis: length of the clean reads (nucleotides). Read counts above the black line are positive sRNA strand reads (sense reads); and read counts below the black line are negative sRNA strand reads (antisense reads).

piRNA abundance is identical for different TE classes in *H. cydno*, *H. melpomene* and F1 inter-specific female hybrids

H. cydno, *H. melpomene* and hybrid female ovaries have equivalent sRNA read length distributions when sRNA reads mapping to all the pathways and TE classes are analysed together. To investigate whether there was a particular class of TE that did not have its correspondent piRNA pool in the hybrids I compared read length distributions of sRNAs that mapped to each different class of TE: 1) LTR; 2) LINE; 3) DNA; 4) RC; 5) SINE; and 6) Unclassified; for both the reference *H. cydno* and *H. melpomene* repeat libraries.

There were very few TEs mapping to the SINE class for all the samples as these elements are not well characterised in the repeat annotations. Regardless, for all the TE classes, there is no evident difference in read length distributions between *H. melpomene*, *H. cydno* and hybrid female ovary sRNAs mapping to: 1) DNA TEs (Supplementary Figure S7-S10); RC TEs (Supplementary Figure S11-S14); LTR TEs (Supplementary Figure S15-S18); LINE TEs (Supplementary Figure S19-S22); SINE TEs (Supplementary Figure S23-S24); or Unclassified TEs (Supplementary Figure S25-S28).

TEs over-expressed in the F1 female hybrids and their corresponding sRNAs

There is no global TE de-repression in the F1 hybrids. However, there are TEs differentially expressed in the hybrids (Figure 3, Table 3, Supplementary S2). I wanted to understand whether the over represented TE transcripts in the hybrids could be explained by the content of the hybrid piRNA pool. Read distributions of sRNAs mapping to these TEs in F1 female hybrids are similar to *H. melpomene* sample sRNA read distributions. There is no significant difference between *H. melpomene* and F1 female sRNA distributions for none of the transposable elements that are is abundant in the hybrids (Figure S29-

S33). *H. cydno* x *melpomene* hybrids have more mRNA reads mapping to these elements but they also have, on average, more sRNAs mapping to the respective TE. The exception is Jockey where *H. melpomene* samples have less mRNAs mapping but slightly more sRNAs (Figure 7).

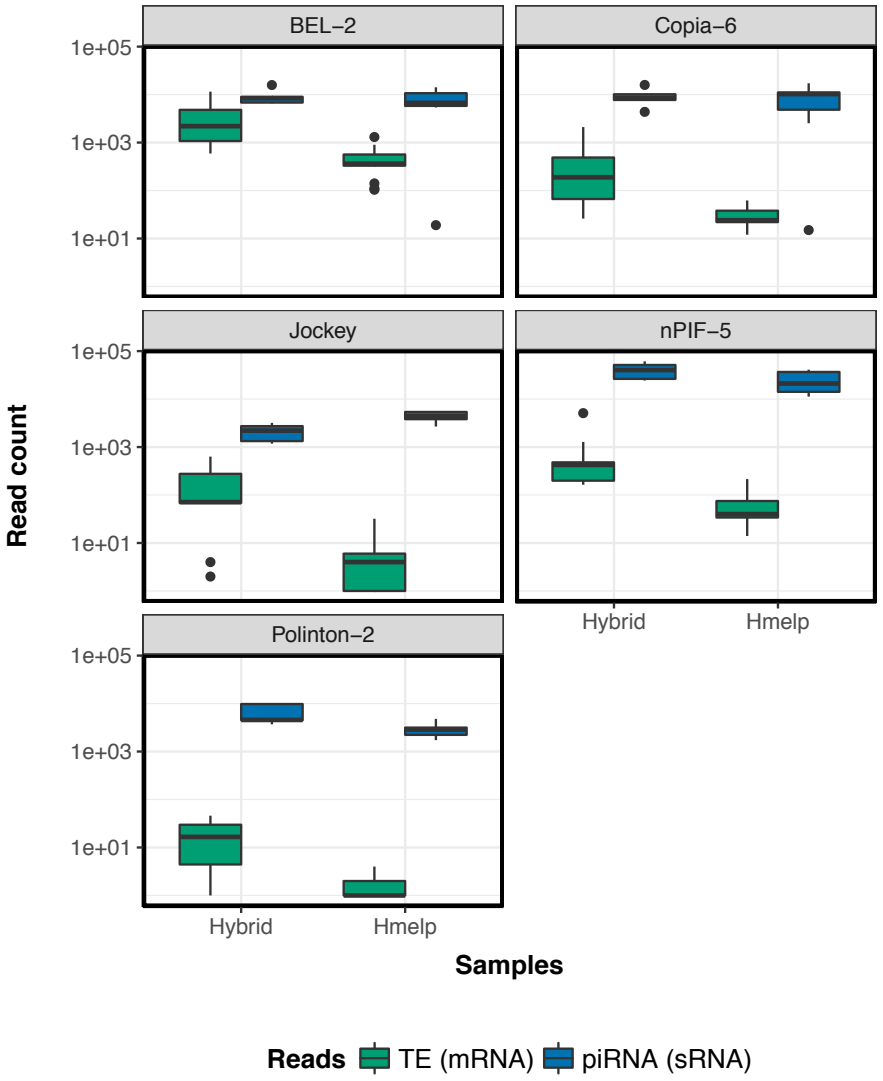


Figure 7. Normalised piRNA and TE read count analysis

There is a positive trend between TE read counts and piRNA read counts for the different samples. Read counts for de-repressed TE in F1 hybrids compared to *H. melpomene* samples: BEL-2, Copia-6, Jockey, nPIF-5 and Polinton-2. Hybrids have larger number of

transcripts for these TEs then *H. melpomene* but they are also have the corresponding piRNAs and these show a read size distribution equivalent to *H. melpomene* samples (Figure S29-S33).

Differentially under-expressed TEs in the hybrids neighbour under-expressed genes

TEs can affect nearby gene activity (Slotkin and Martienssen, 2007; Wang *et al.*, 2013). I examined whether under-expressed TEs are positively correlated with variation in gene expression – if a TE was under-expressed are genes in its vicinity also repressed? If so, suppression of adjacent protein coding genes by TEs could be correlated to the sterility phenotype. Within 2 kb of the 9 under-expressed TEs there are 11 protein coding genes. All under-expressed TEs are also flanked by under-expressed genes (Figure 8, Table 4). One of these genes, HMEL041988, overlaps the sterility QTL discussed in Chapter 3, Sterility in *Heliconius cydno* x *Heliconius melpomene* F1 female hybrids: a phenotypic and gene expression study of hybrid incompatibilities. 5 out of the 11 genes that are flanked by under-expressed TEs could got get transferred from the *H. melpomene* reference annotation to the *H. cydno* reference annotation (Chapter 3). Protein coding genes flanked by TEs have expression patterns that are more similar between *H. cydno* and F1 hybrids than between *H. melpomene* and the hybrids (Figure 8, Table 4).

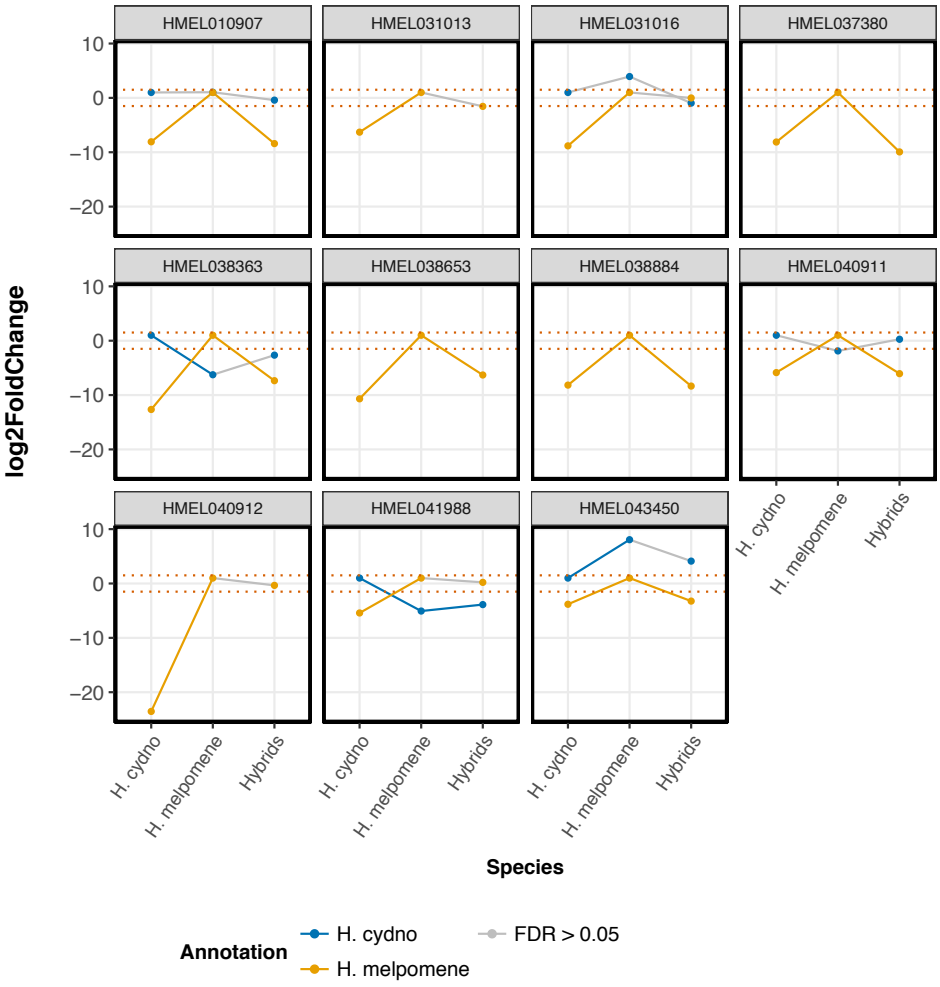


Figure 8. Patterns of expression for genes within 2 kb of under-expressed TEs

Gene expression from the analysis performed on the *H. melpomene* reference genome and annotation (yellow) and on the *H. cydno* reference genome and annotation (blue). Genes are within 2 kb of under-expressed TEs. Points represent a gene in each of the three different sample groups. Expression for each gene calculated with *H. melpomene* (yellow) or *H. cydno* (blue) samples as the baseline. Negative log2 fold values – gene expression is lower than the baseline. Positive log2 fold values – gene expression is greater than the baseline. Gene expression values of the three different groups (*H. cydno*, *H. melpomene* and hybrids) are linked by yellow, blue or grey

lines. Yellow and blue lines represent significant results (FDR < 0.05 and log2 fold change > |1.5|), grey lines represent non-significant results (FDR > 0.05 and/or log2 fold change < |1.5|). Dotted red lines delineated the |1.5| log2 fold change significance threshold. Genes HMEL031013, HMEL037380, HMEL038653, HMEL038884 and HMEL040912 could not be transferred to the *H. cydno* annotation.

Scaff.	Gene	log2 FC. gene	TE	log2 FC. TE	Protein family / Domain
213052	010907	-5.78	Gyp-6	-8.41	BEIGE/BEACH related
216006	037380	-2.39	Gyp-10	-9.94	NA
217014	038363	-2.39	Gyp-10	-7.36	NA
217020	038653	-2.39	Gyp-10	-6.31	Reverse transcriptase catalytic domain
218001	038884	-2.39	Gyp-10	-8.34	GAG/POL/ENV polyprotein
220005	040911	-2.39	Gyp-10	-6.06	Regulator of G protein signalling domain
220005	040912	-2.39	Gyp-10	-0.34	Uncharacterised protein
221012	041988	-2.39	LTR-10	0.21	NA
203069	043450	-4.35	Mariner-N29	-3.24	Zinc finger C2H2 type domain, Sodium solute symporter family
209007	031013	0.39	LTR-10	-1.56	NA
209007	031016	-5.45	LTR-10	-0.01	NA

Table 4. Genes within 2 kb of under-expressed TEs

Gene name and scaffold position within 2 kb of under-expressed TEs. Log2 Fold changes of gene and TE in the hybrids as calculated using DEseq2. Protein family/domain predicted with InterProScan. All scaffold names start with Hmel; gene names start with HMEL.

Discussion

When two genomes come together in hybrids it is common to observe disruption of the genome and transcriptome. Specifically, piRNA deficiencies and TE mobilization have been extensively linked to hybrid sterility in inter- and intra-specific *Drosophila* crosses (Brennecke *et al.*, 2008; Kelleher *et al.*, 2012). For example, in *D. melanogaster* x *D. simulans* artificial rescue hybrids, exhibit a global de-regulation of TEs. In contrast, interspecific hybridization between *H. melpomene* males and *H. cydno* females does not result in F1 with significant changes in TE expression.

The global de-repression observed in *Drosophila* hybrids has been linked to deficient piRNA production which is seen as a shift in the ovary piRNA length distribution in hybrids from 23-30 nucleotides to 18-22 nucleotides (Kelleher *et al.*, 2012). In contrast, F1 *H. melpomene* x *H. cydno* female piRNA length distributions are the same in the hybrids and the parental species. piRNA length distributions in *Heliconius* F1s are analogous to observations made for *D. buzzatii* x *D. koepferae* hybrids, which also do not have a deficient global piRNA production. However, F1 *D. buzzatii* x *D. koepferae* hybrids have 15.2% of the expressed TE families de-regulated in F1 hybrid ovaries (Romero-Soriano et al 2017). In summary, both the fact that there is no difference in piRNA gene expression between the *H. melpomene*, *H. cydno* and F1 hybrids; and also no difference in piRNA length distribution indicates that the piRNA pathway is functional in the hybrids. This disproves the piRNA global failure hypothesis in *H. cydno* x *H. melpomene* female hybrids.

In *H. cydno* x *H. melpomene* females Copia-6, Jockey and Polinton-2 have fold changes significantly different between the parents and are de-regulated in hybrids for but there is no significant different between the sRNA distributions mapping to these elements disproving the maternal cytotype hypothesis. Moreover, in hybrid dygenesis, an overall increased in recombination rates is observed (Kidwell *et al.*, 1977; Kidwell, 1983; Hill *et al.*, 2016). Between *H. melpomene*, *H. cydno* and the hybrids there are no

genomic rearrangements and the rates of recombination are equivalent similar across the genome (Davey *et al.,* 2017). Sequence divergence between maternal piRNAs and paternal TE transcripts may lead to decreased efficacy of silencing in *H. cydno* x *melpomene* hybrids and, consequently, TE over-expression. However, the presence of under-expressed TEs in the hybrids makes this hypothesis less plausible.

TE over-expression in hybrid genomes has been observed in plants and animals. Generally TE over-expression is considered to be the common outcome following hybridization (Kawakami *et al.*, 2011; Kelleher *et al.*, 2012; Hill *et al.*, 2016). However, whole-genome studies have reported TE under-expression in hybrids. For example, in lake whitefish hybrids, over 1/3 of TE are under-expressed; and in sunflowers, F1s have most TEs under-expressed (Dion-Côté *et al.*, 2014; Renaut *et al.*, 2014). Repression of TEs in the hybrids could follow the loss of epigenetic silencing leading to reinforcement of TE silencing in *trans*. This phenomenon and its consequences has been largely ignored, contrasting to over-expression (Rigal *et al.*, 2016). However, in *H. cydno* x *H. melpomene* hybrids I only observed a few under- and over-expressed TEs. Both genic expression, repeat element abundance and non-coding RNA abundance seem to be more similar between *H. cydno* and the hybrids than between *H. melpomene* and the hybrids. As piRNAs are maternally inherited and *H. cydno* is the mother of these crosses this pattern is not unexpected. However, the *H. cydno* annotation is less complete both for genetic features and in repeat element content and so I cannot exclude that the observed similarity might be an artefact of the annotation.

Reproductive isolation can arise from divergent molecular evolution between populations and it considered a step towards speciation. This is the first time a study involving sRNA and mRNA sequencing from several samples has been conducted to investigate the possible link between piRNA pathway mediated silencing and hybrid sterility, in any insect outside *Drosophila*. As in many cases, evolutionary patterns seen in *Drosophila* do not necessarily apply to other insects (e.g. Lewis *et al.* (2017). Here, I have shown that the sterility

phenotype observed in F1 female hybrids does not correlate with either piRNA pathway coding gene expression differences, or TE and piRNA abundance disproving both the maternal cytotype hypothesis and the piRNA global failure hypothesis. All my results suggest that the female germline is successfully protected against TE mobilization and this is not the cause of hybrid sterility.

Here I focused on hybrid female sterility from a hybrid cross between a *H. cydno* female and a *H. melpomene* male. While, for completeness, it would be interesting to perform the experiment using the opposite crossing scheme, the reciprocal cross (*H. melpomene* female x *H. cydno* male) has never been successful (details in Naisbit *et al.,* 2002). In a fashion similar to what it observed between *D. melanogaster* and *D. simulans* (Carracedo *et al.* 1998), between *H. cydno* and *H. melpomene*, we have only been able to perform hybrid crosses in one direction. Despite my inability to perform the reciprocal cross, the robustness of the results for the direction reported here are not affected. I have set out to investigate whether hybrid female infertility of a cross between *H. cydno* female and *H. melpomene* male might be corrected to TEs and/or sRNA differences between *H. cydno* and *H. melpomene* and concluded that does not seem to be the case.

In the future, identifying other piRNA genes by sequence homology with arthropod databases and calculating how fast they are evolving in comparison to siRNA or miRNA genes will further clarify why the piRNA pathway is functional in the *H. cydno* x *H. melpomene* hybrids. Moreover, there seems to be some evidence that TE down-regulation affects expression of nearby genes in a fashion similar to what has been identified in plants. I need to explore this by quantifying if expression of genes flanking TEs diminishes with TE proximity. If this is the case and there is a negative correlation between the distance to a down-regulated TE and gene expression then the sterility phenotype could be due to gene mis-expression as an indirect consequence of the piRNA pathway silencing. Finally, transcriptional silencing via methylation could instead be the cause of the observed de-regulated TEs. To

test this hypothesis we will be analysing methylation polymorphisms and patterns of inheritance between the fertile *H. melpomene* father and *H. cydno* mother and the infertile hybrid progeny. Lastly, it may also be worth investigating the expression of micro RNAs through the integrated analysis of micro RNA and mRNA expression in a fashion similar to what has been done for piRNAs as there is the possibility that it is a deregulation of a micro RNA that is behind the observed phenotype.

Supplementary Tables

Supplementary Table S1.

Sample name	Species	Tissue	Stage	Raw Reads	Clean Reads	Error Rate(%)	Q20(%)	GC Content(%)
AP23	H. cydno	Ovary	Young	26837823	25982919	0.02	97.38	41.01
AP28	H. cydno	Ovary	Young	39138123	38706191	0.01	98.61	44.52
AP63	H. cydno	Ovary	Young	33777712	32704267	0.01	97.75	40.83
AP67	H. cydno	Ovary	Young	37359340	36752670	0.01	98.47	41.14
AP21	H. cydno	Ovary	Young	32244551	31192828	0.01	97.82	41.56
AP20	H. cydno	Ovary	Young	33632548	32592198	0.02	97.6	40.61
AP19	H. cydno	Ovary	Young	31459005	30474372	0.02	97.51	40.34
AP35	H. melpomene	Ovary	Young	35018481	34645187	0.01	98.53	41.08
AP94	H. melpomene	Ovary	Young	39903075	39134431	0.01	98.16	41.51
AP34	H. melpomene	Ovary	Young	27811550	27296213	0.01	98.2	41.42
AP37	H. melpomene	Ovary	Young	33410348	32682152	0.01	98.49	42.15
AP71	H. melpomene	Ovary	Young	34856038	34487192	0.01	98.42	41.36
AP88	H. melpomene	Ovary	Young	38486006	38121619	0.01	98.53	42.71
AP93	H. melpomene	Ovary	Young	31497198	30839548	0.01	98.38	41.24
AP55	H. melpomene	Ovary	Mature	37934077	37335336	0.01	98.17	39.76
AP77	H. melpomene	Ovary	Mature	34322656	33261770	0.01	98.28	40.46
AP80	H. melpomene	Ovary	Mature	36157750	35149502	0.01	97.97	40.28
AP89	H. melpomene	Ovary	Mature	34318423	33383486	0.01	98.04	41.51
AP141	H. melpomene	Ovary	Mature	33844256	32934318	0.01	97.88	39.58

Supplementary Table S1 (cont).

Sample name	Species	Tissue	Stage	Raw Reads	Clean Reads	Error Rate(%)	Q20(%)	GC Content(%)
AP142	H. melpomene	Ovary	Mature	35328097	34348031	0.01	98	39.69
AP54	H. cydno X H melpomene	Ovary	Young	36589398	36249200	0.01	98.42	42.61
AP38	H. cydno X H melpomene	Ovary	Young	35305269	34283164	0.01	98.19	40.41
AP39	H. cydno X H melpomene	Ovary	Young	39788204	39104796	0.01	98.42	41.2
AP53	H. cydno X H melpomene	Ovary	Young	34638287	34288204	0.01	98.33	43.8
AP58	H. cydno X H melpomene	Ovary	Young	44218820	43693866	0.01	98.4	41.98
AP66	H. cydno X H melpomene	Ovary	Young	40503519	39816955	0.01	98.5	41.23
AP65	H. cydno X H melpomene	Ovary	Young	40575350	39711736	0.01	98.47	42.93
AP41	H. cydno X H melpomene	Ovary	Young	36564414	35536413	0.01	98.37	40.5
AP52	H. cydno X H melpomene	Ovary	Young	26051783	24931600	0.01	98.51	43.27
AP70	H. cydno X H melpomene	Ovary	Young	24514381	23726336	0.01	98.81	41.2

Supplementary Table S1. mRNA sample sequencing summary statistics

Sample ID, species, tissue, stage of collection for mRNA 150bp PE directionally sequenced reads. Sequencing summary statistics presented as total number of reads sequenced – Raw reads; reads left in sample after quality filter – Clean reads. Error rate, Q20 and GC content statistics also calculated.

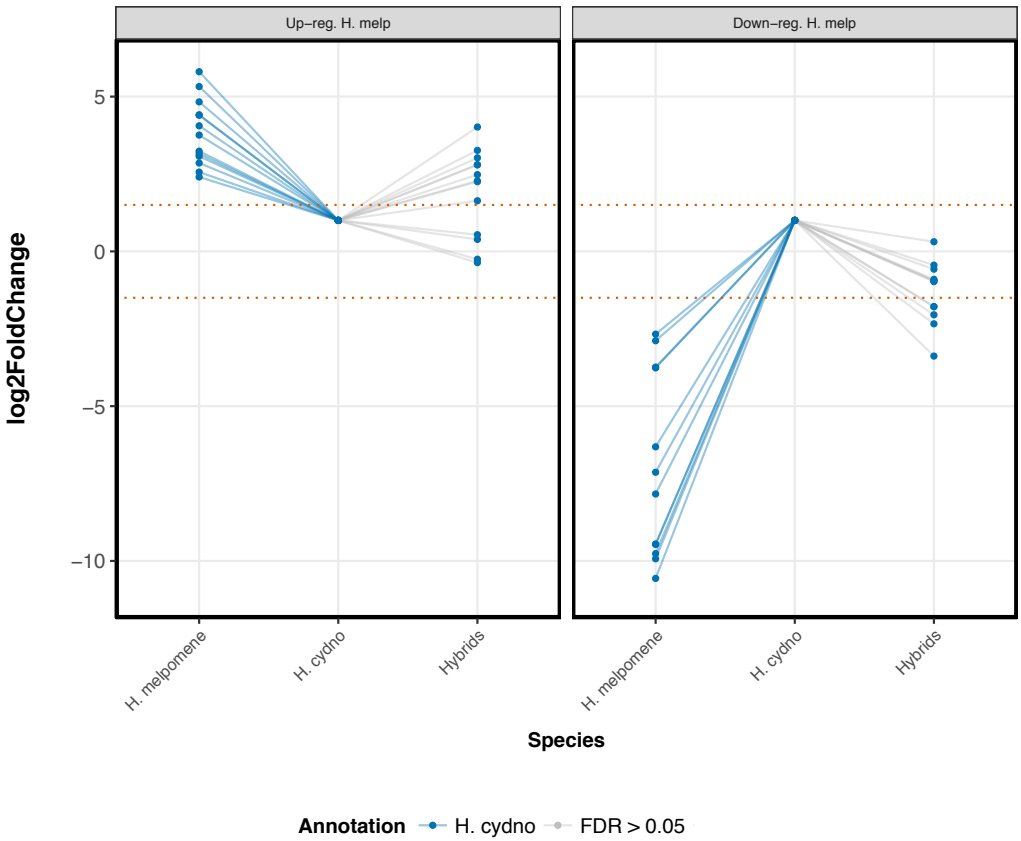
Supplementary Table S2.

Sample name	Species	Tissue	Raw reads	Clean reads
AP2	*H. melpomene*	Ovary	46726424	37664842
AP4	*H. melpomene*	Ovary	40439330	49678301
AP5	*H. cydno*	Ovary	47160526	41829618
AP6	*H. cydno*	Ovary	46691883	48073178
AP10	*H. cydno*	Ovary	48036776	53015118
AP13	*H. melpomene*	Ovary	49401953	51006892
AP16	*H. melpomene*	Ovary	43350525	44678823
AP17	*H. cydno*	Ovary	39411079	47225286
AP22	*H. melpomene*	Ovary	41235484	48265998
AP30	*H. melpomene*	Ovary	50055614	47861217
AP33	*H. melpomene*	Ovary	45017899	42723367
AP50	*H. melpomene* X *H. cydno*	Ovary	36718688	50233709
AP57	*H. melpomene* X *H. cydno*	Ovary	46337008	45726693
AP59	*H. melpomene* X *H. cydno*	Ovary	38895269	47546180
AP60	*H. melpomene* X *H. cydno*	Ovaries	51797924	39516993
AP72	*H. melpomene* X *H. cydno*	Ovaries	47008978	40400948

Supplementary Table S2. sRNA sample sequencing summary statistics

Sample ID, species and tissue for sRNA SE 50bp sequenced reads. Sequencing summary statistics presented as total number of reads sequenced – Raw reads; reads left in sample after quality filter – Clean reads.

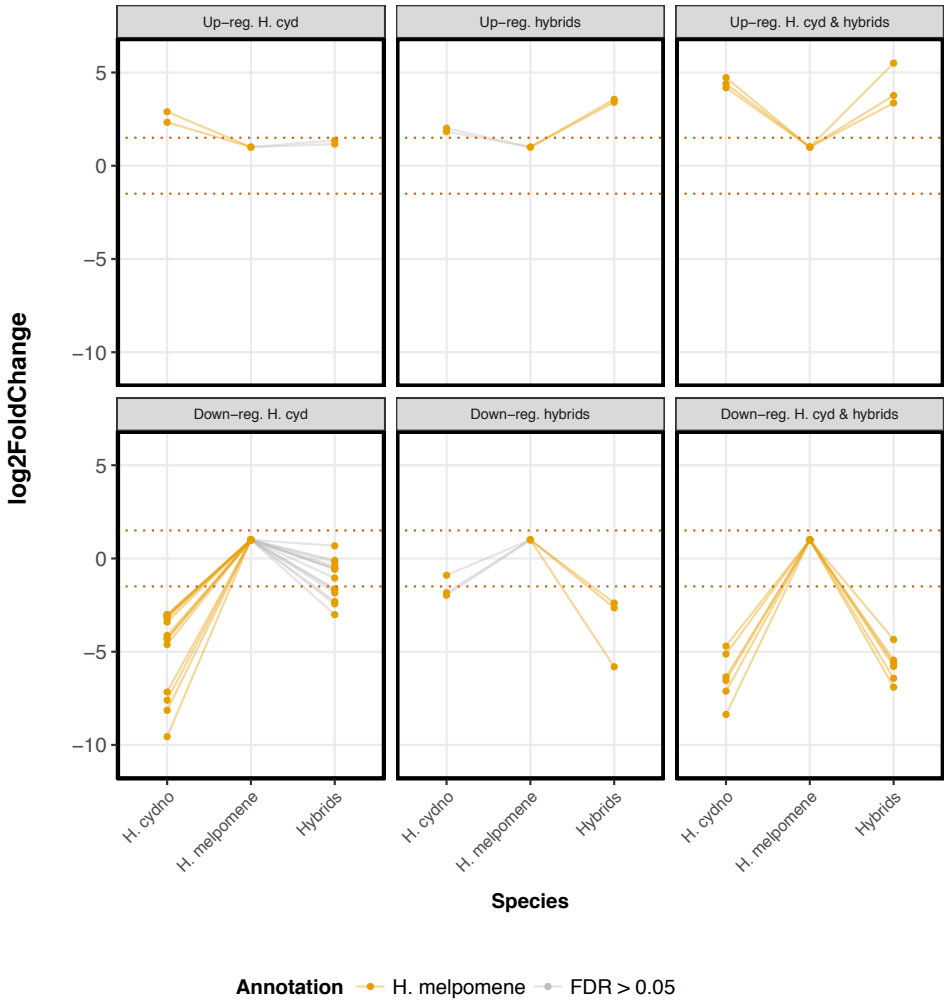
Supplementary Figures



Supplementary Figure S1. Patterns of expression for the differential expressed TEs identified with the reference *H. cydno* repeat annotation

TEs differentially expressed identified with the analysis performed on the *H. cydno* reference repeat annotation. Points represent a TE in each of the three different sample groups. Expression for each TE calculated with *H. cydno* samples as the baseline. Negative log2 fold values – TE expression is lower than in *H. cydno*. Positive log2 fold values – TE expression is greater than in *H. cydno*. Differentially expressed TEs are separated by their expression patterns: 1) up-

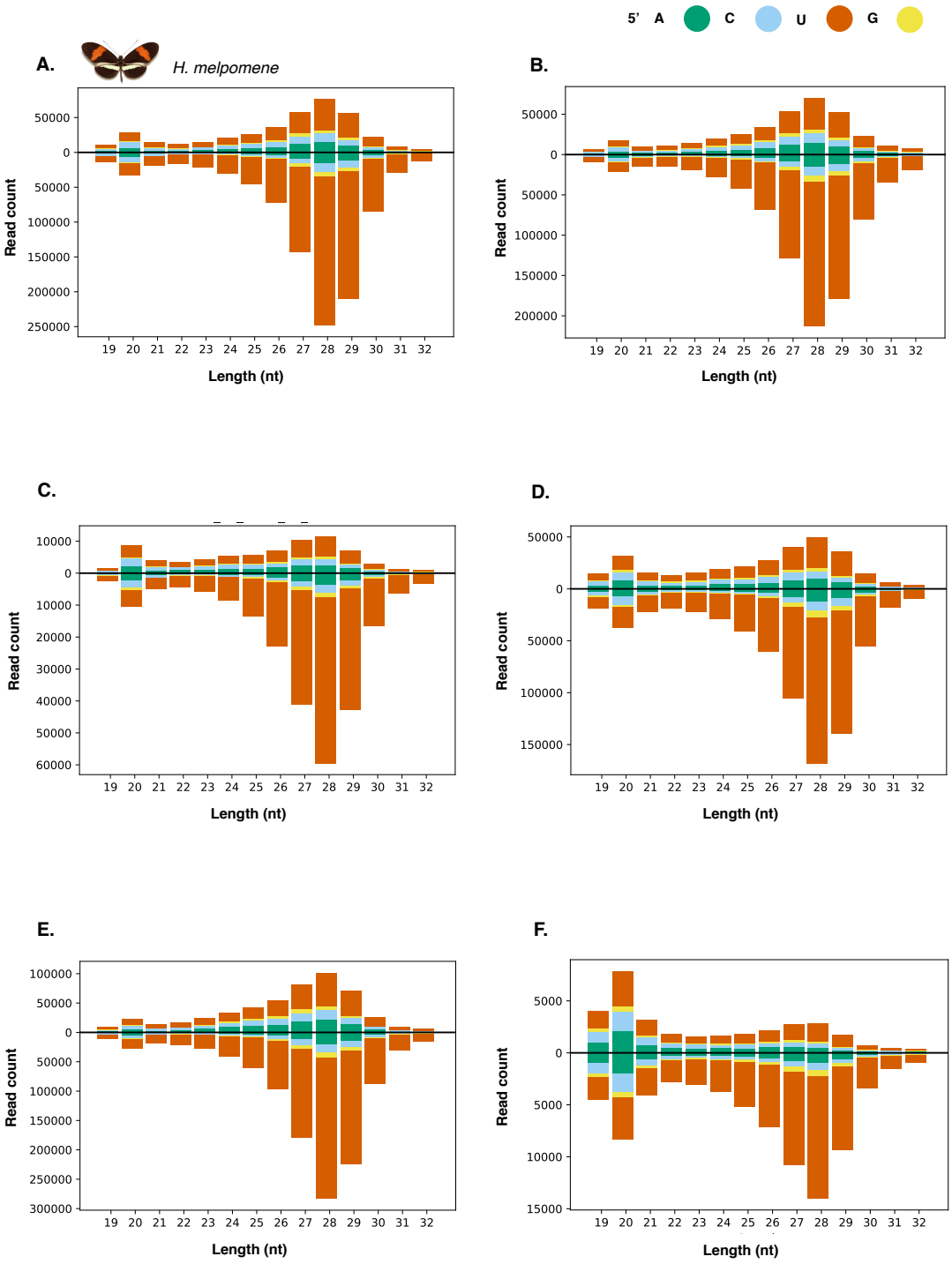
regulated in *H. melpomene*, 2) down-regulated in *H. melpomene*. TE expression values of the three different groups (*H. cydno*, *H. melpomene* and hybrids) are linked by blue or grey lines. Blue lines represent significant results (FDR < 0.05 and log2 fold change > |1.5|), grey lines represent non-significant results (FDR > 0.05 and/or log2 fold change < |1.5|). Dotted red lines delineated the |1.5| log2 fold change significance threshold.



Supplementary Figure S2. Patterns of expression for the differential expressed TEs identified with the reference *H. melpomene* repeat annotation

TEs differentially expressed identified with the analysis performed on the *H. melpomene* reference repeat annotation. Points represent a TE in each of the three different sample groups. Expression for each TE calculated with *H. melpomene* samples as the baseline. Negative log2 fold values – TE expression is lower than in *H. melpomene*. Positive log2 fold values – TE expression is greater than in *H. melpomene*. Differentially expressed TEs are separated by their expression patterns: 1) up-regulated in *H. cydno*, 2) up-regulated in the hybrids, 3)

up-regulated in *H. cydno* and in the hybrids, 4) down-regulated in *H. cydno*, 5) down-regulated in the hybrids, 6) down regulated in *H. cydno* and in the hybrids. TE expression values of the three different groups (*H. cydno*, *H. melpomene* and hybrids) are linked by yellow or grey lines. Yellow lines represent significant results (FDR < 0.05 and log2 fold change > |1.5|), grey lines represent non-significant results (FDR > 0.05 and/or log2 fold change < |1.5|). Dotted red lines delineated the |1.5| log2 fold change significance threshold.


5' A C U G
A.
H. melpomene
B.
C.
D.
E.
F.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

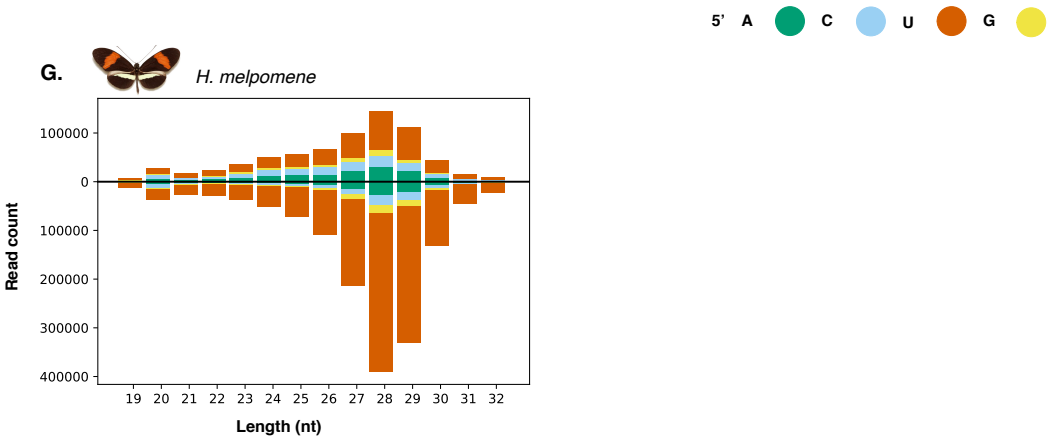


Figure S3. sRNA pool for *H. melpomene*

sRNA read distribution for *H. melpomene* samples mapped to all classes of *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33.

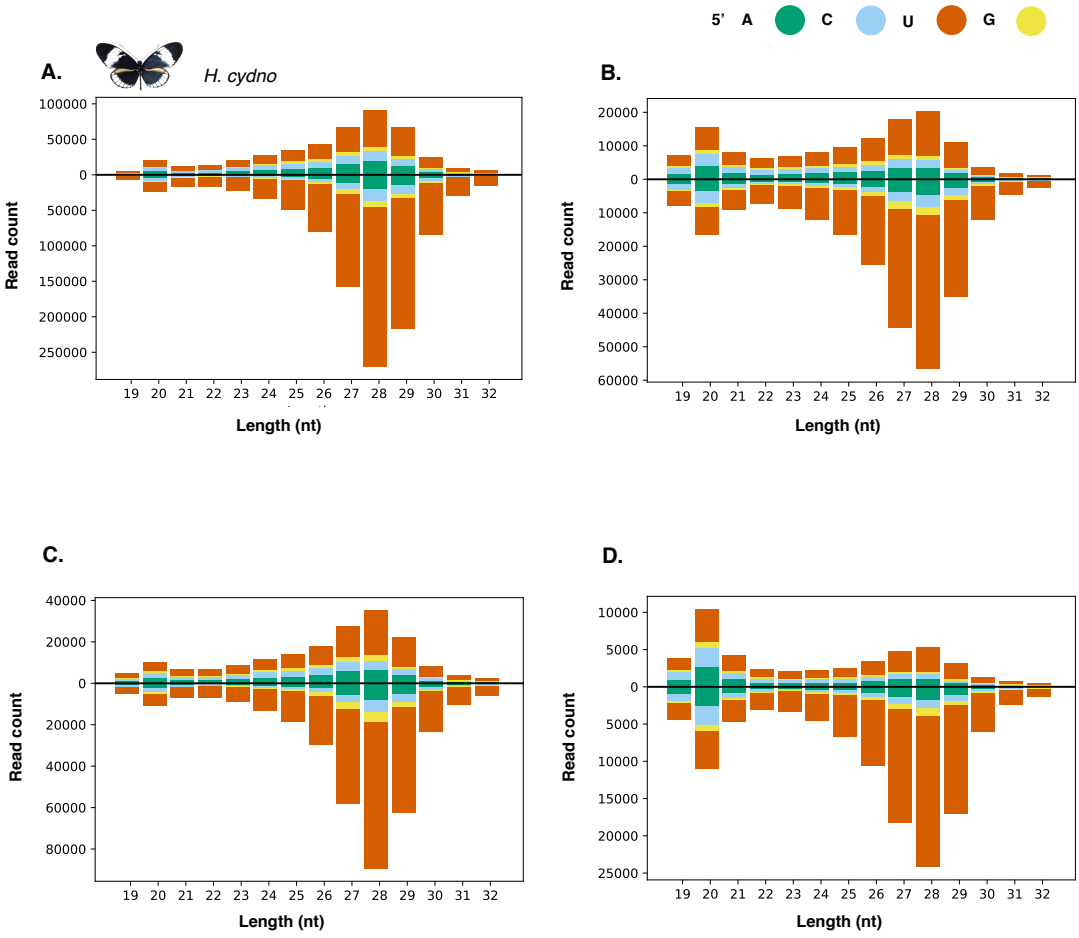


Figure S4. sRNA pool for *H. cydno*

sRNA read distribution for *H. cydno* samples mapped to all classes of *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP5, **B.** Sample AP6, **C.** Sample AP10, and **D.** Sample AP17.

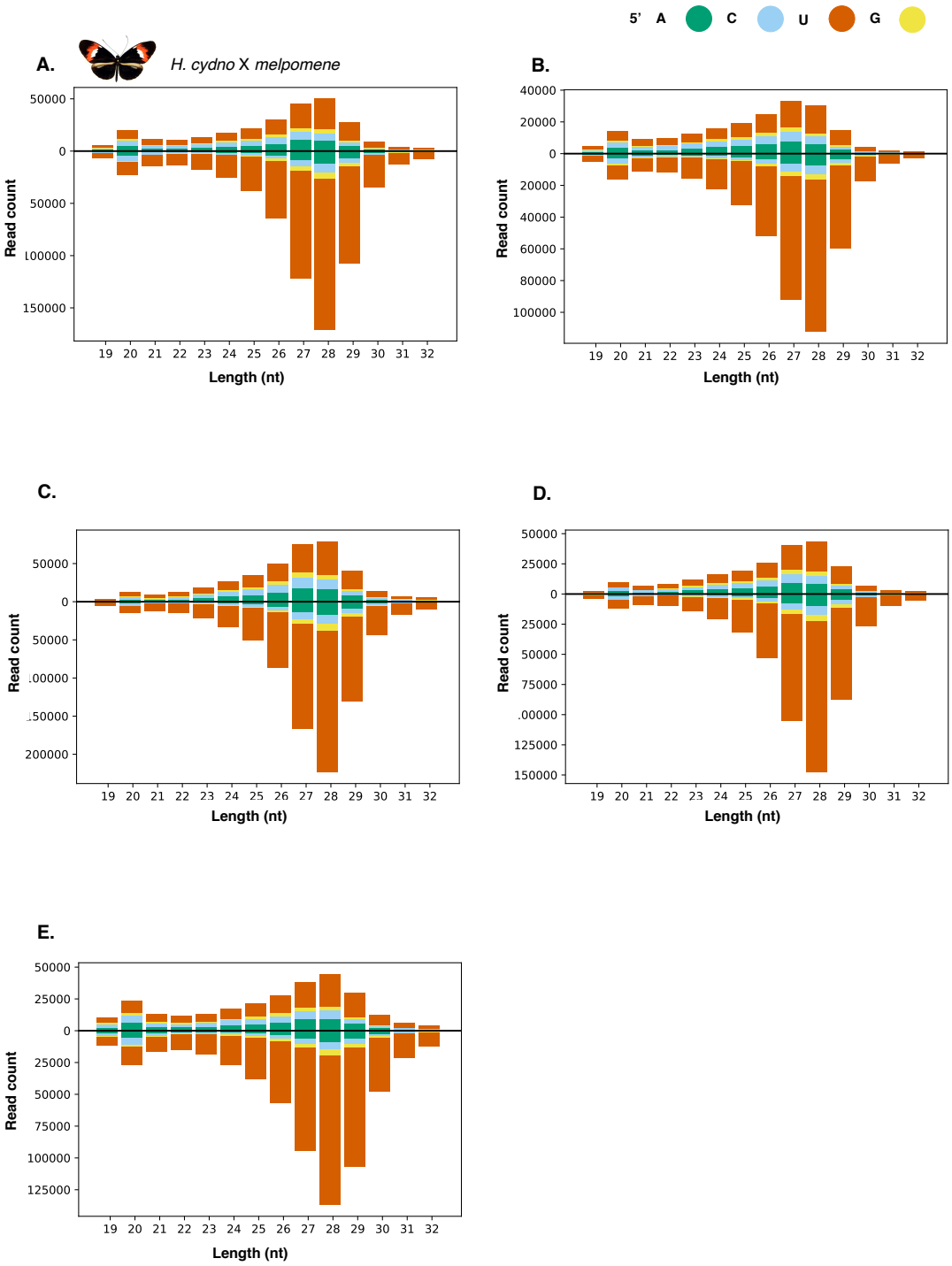

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S5. sRNA pool for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to all classes of *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.

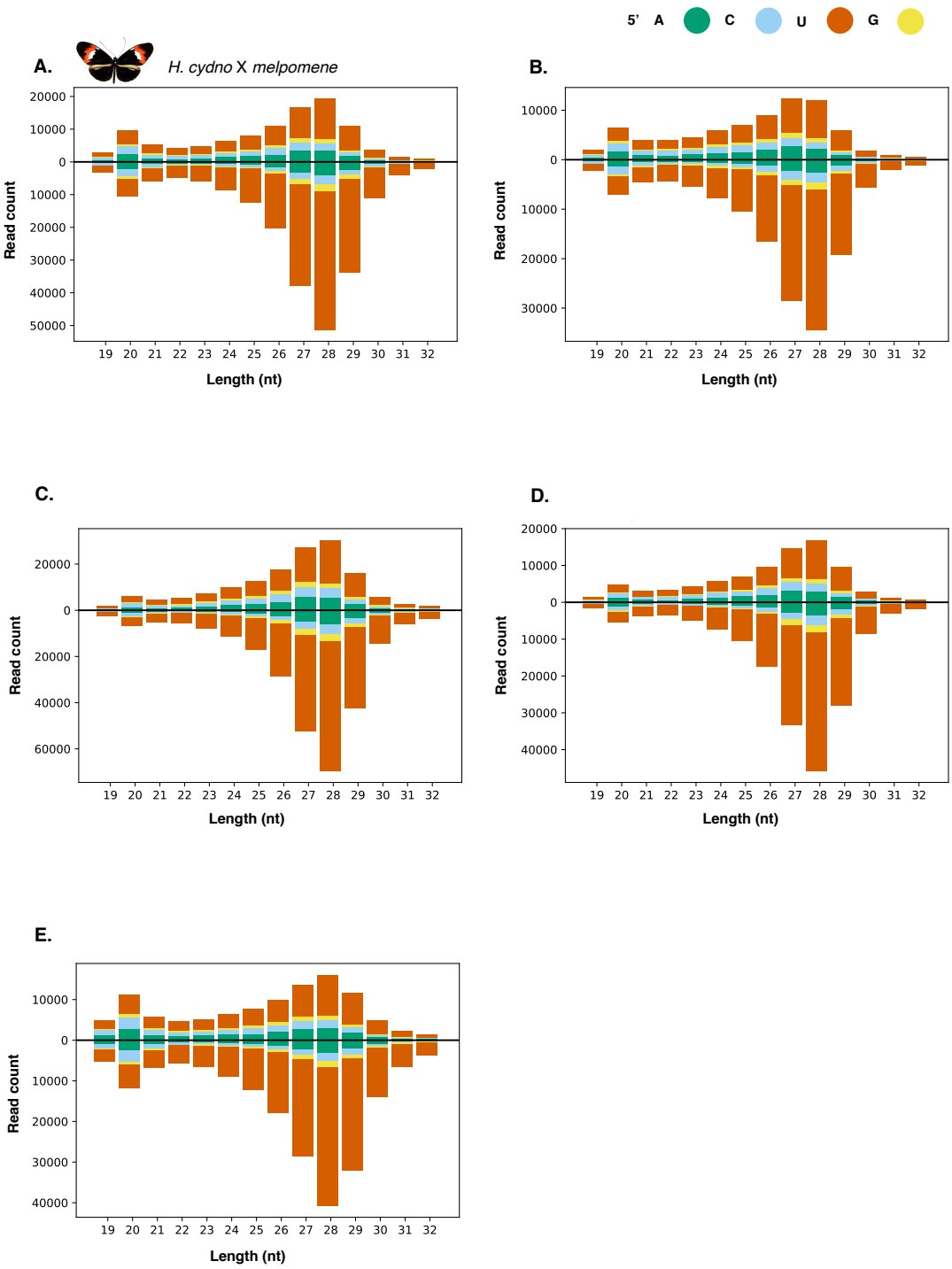
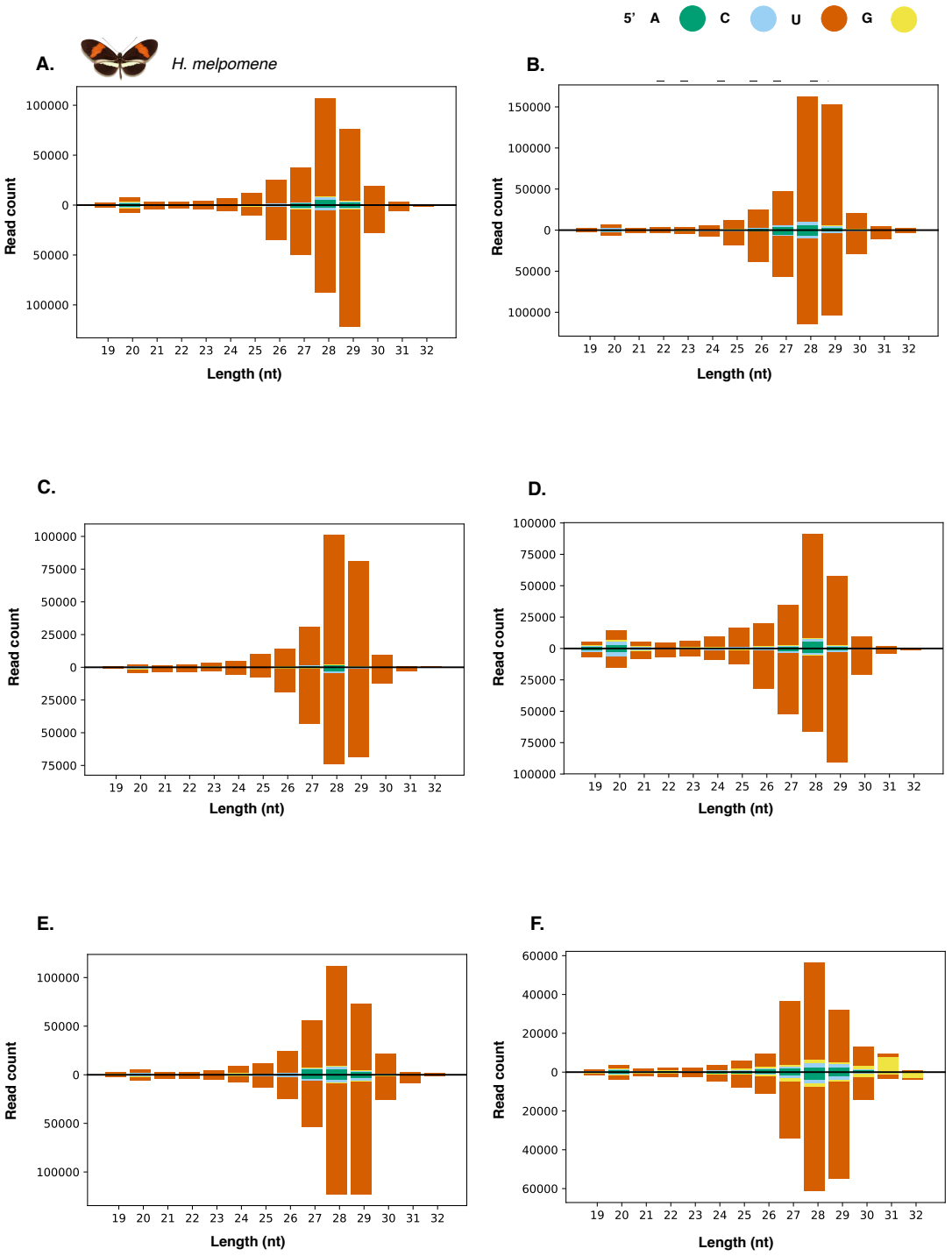

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S6. sRNA pool for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to all classes of *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.


5' A C U G
A.
H. melpomene
B.
C.
D.
E.
F.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

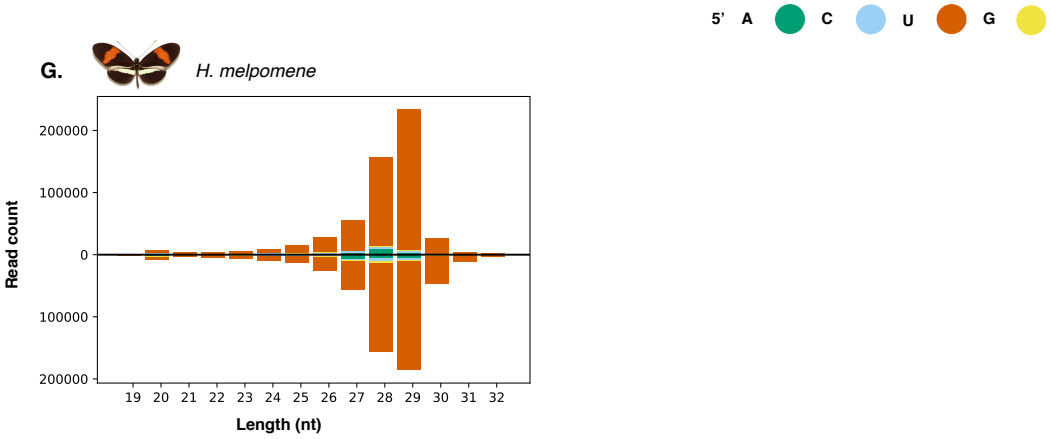


Figure S7. sRNAs mapping to DNA TEs for *H. melpomene*

sRNA read distribution for *H. melpomene* samples mapped to DNA *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33.

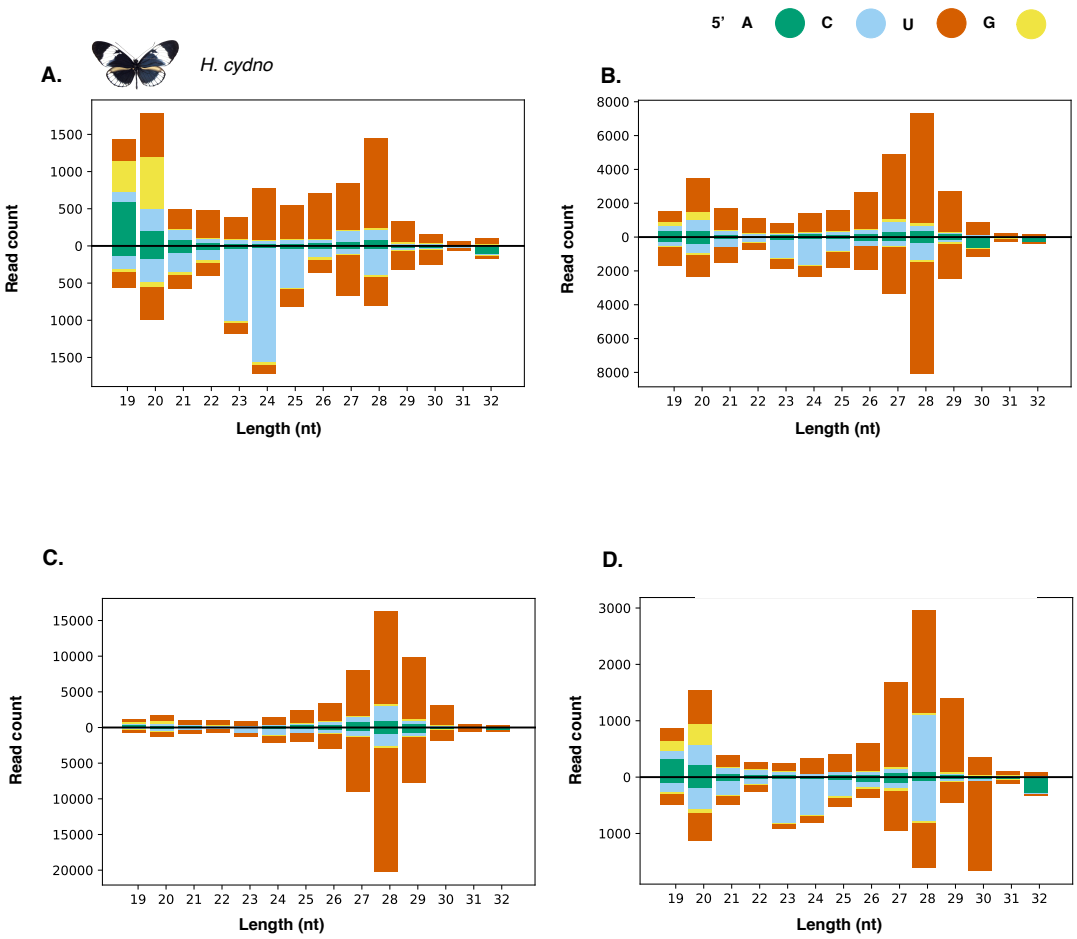


Figure S8. sRNAs mapping to DNA TEs for *H. cydno*

sRNA read distribution for *H. cydno* samples mapped to DNA *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP5, **B.** Sample AP6, **C.** Sample AP10, **D.** Sample AP17.

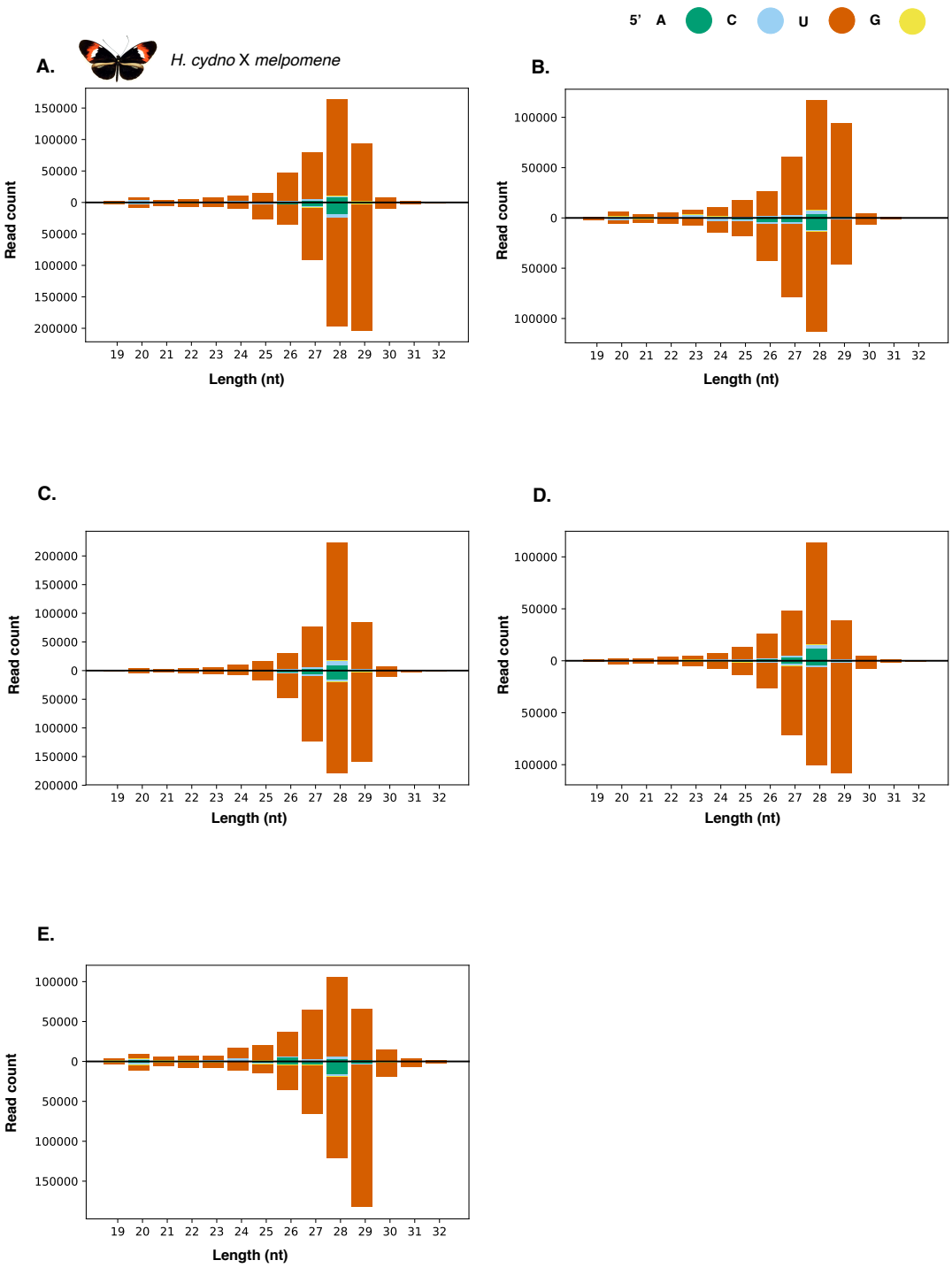

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S9. sRNAs mapping to DNA TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to DNA *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.

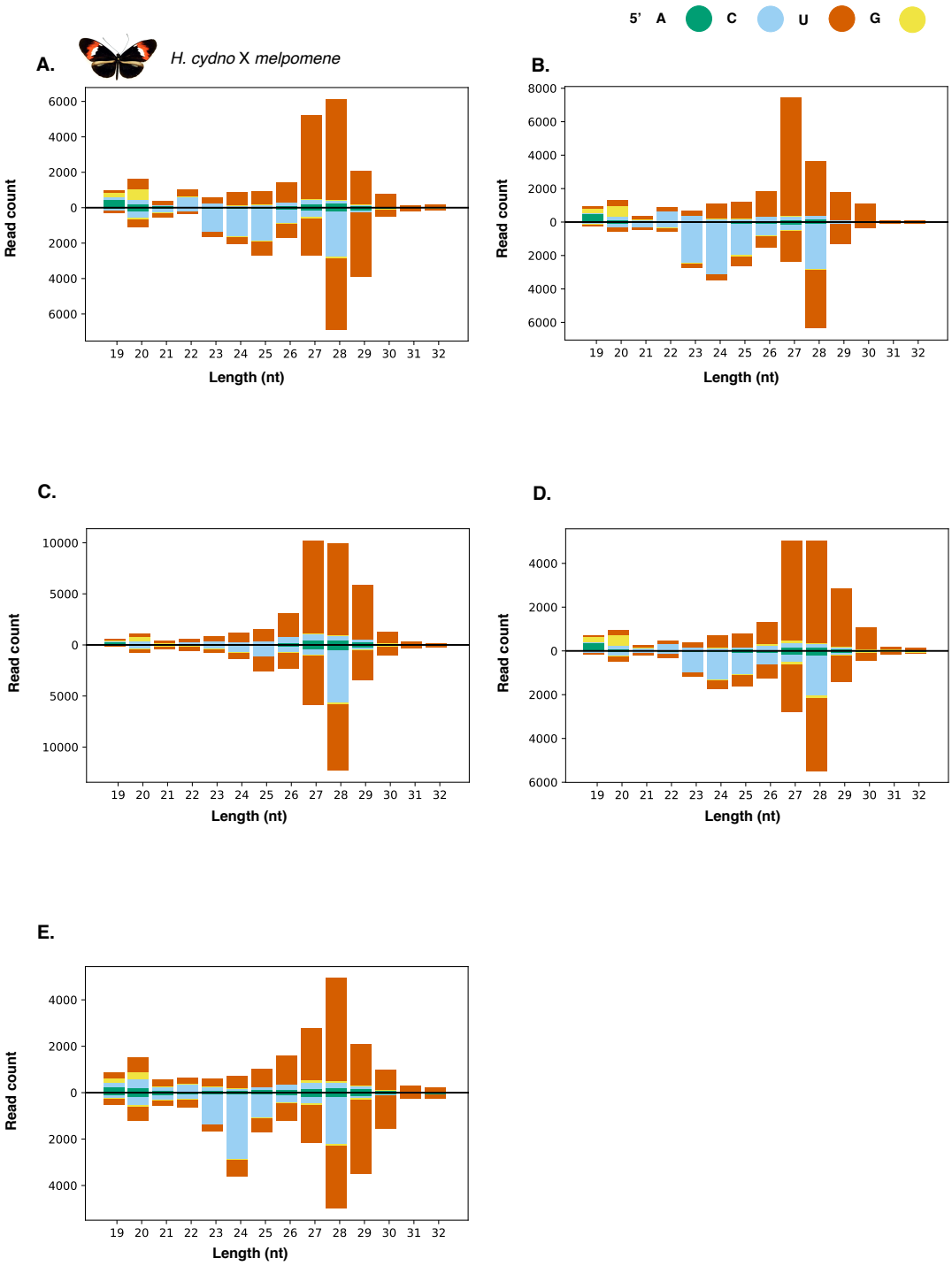
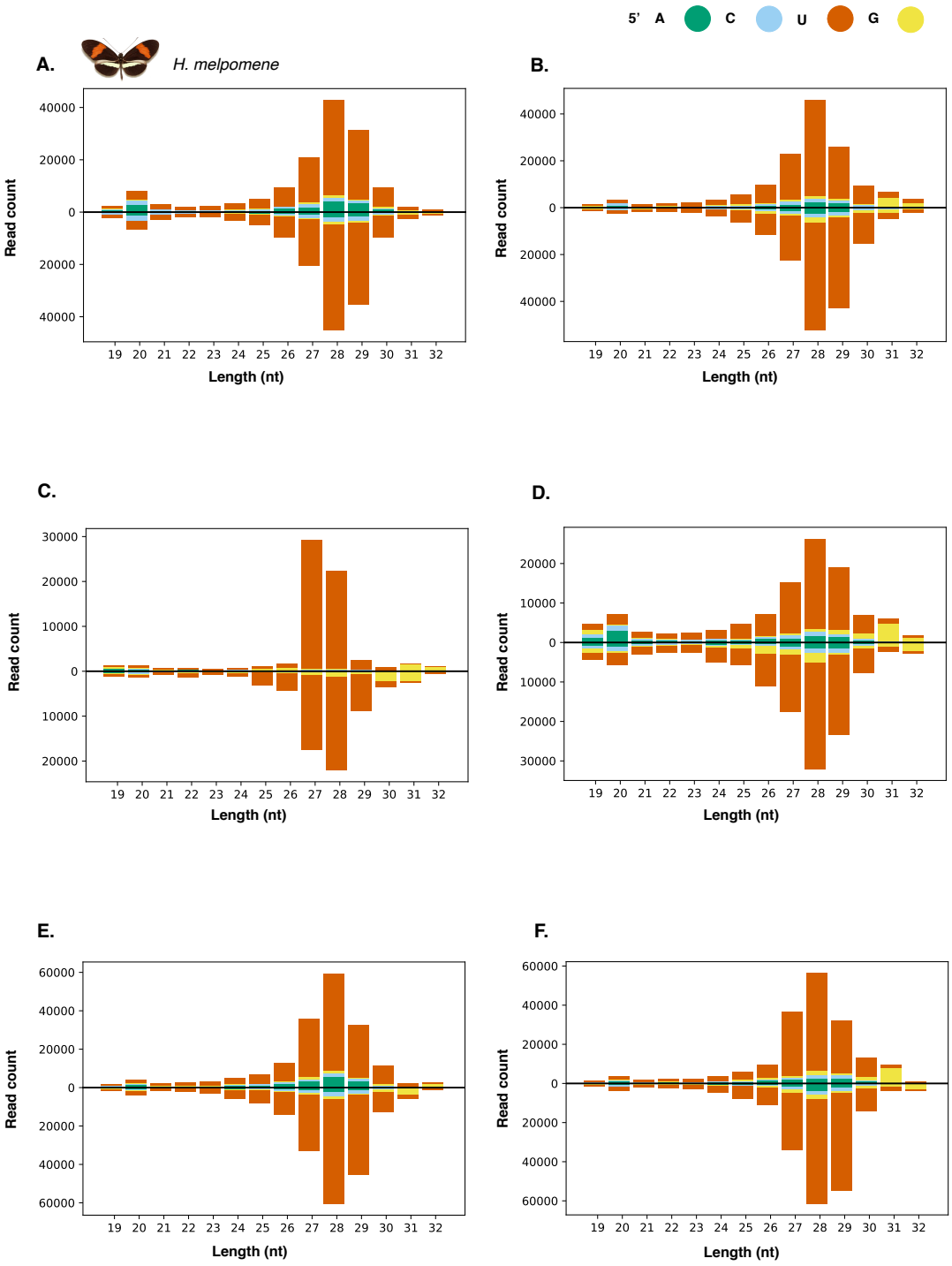

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S10. sRNAs mapping to DNA TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to DNA *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.


5' A C U G
A.
H. melpomene
B.
C.
D.
E.
F.
Read count
Length (nt)
40000
20000
0
20000
40000
19 20 21 22 23 24 25 26 27 28 29 30 31 32
30000
20000
10000
0
10000
20000
60000

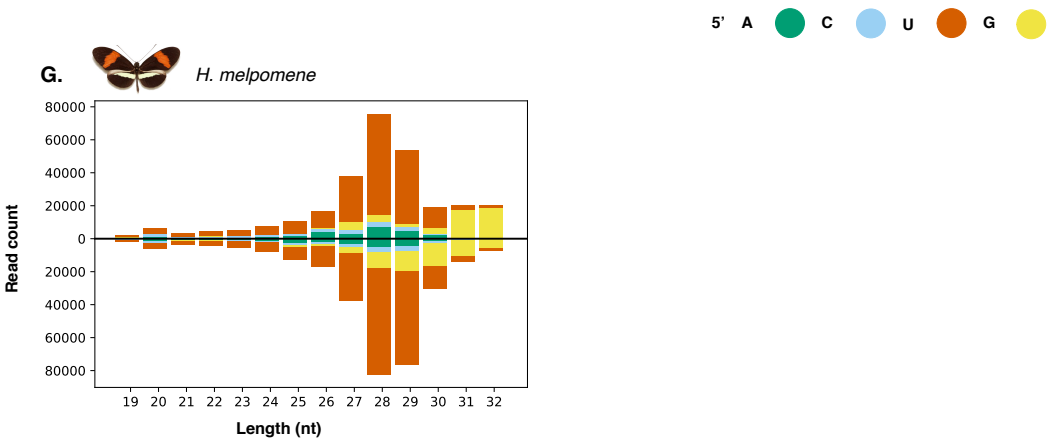


Figure S11. sRNAs mapping to RC TEs for *H. melpomene*

sRNA read distribution for *H. melpomene* samples mapped to RC *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33.

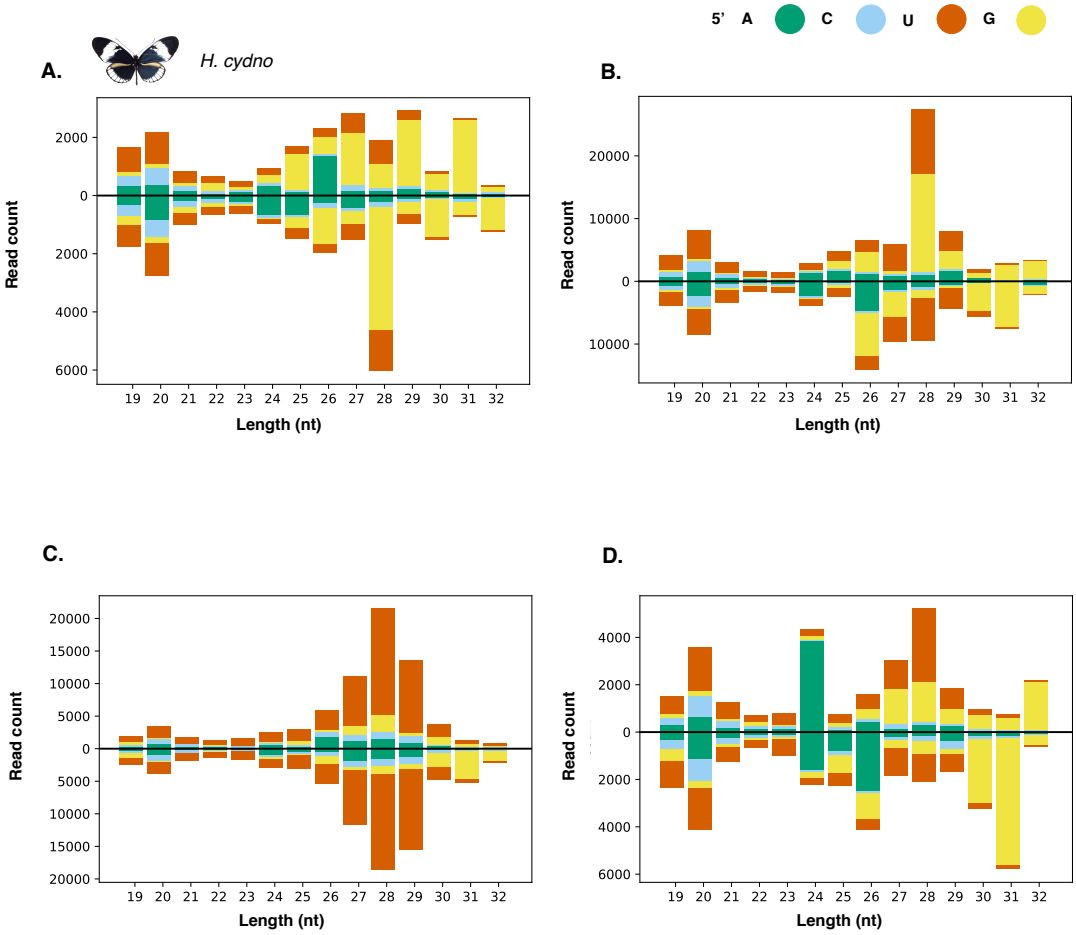


Figure S12. sRNAs mapping to RC TEs for *H. cydno*

sRNA read distribution for *H. cydno* samples mapped to RC *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP5, **B.** Sample AP6, **C.** Sample AP10, **D.** Sample AP17.

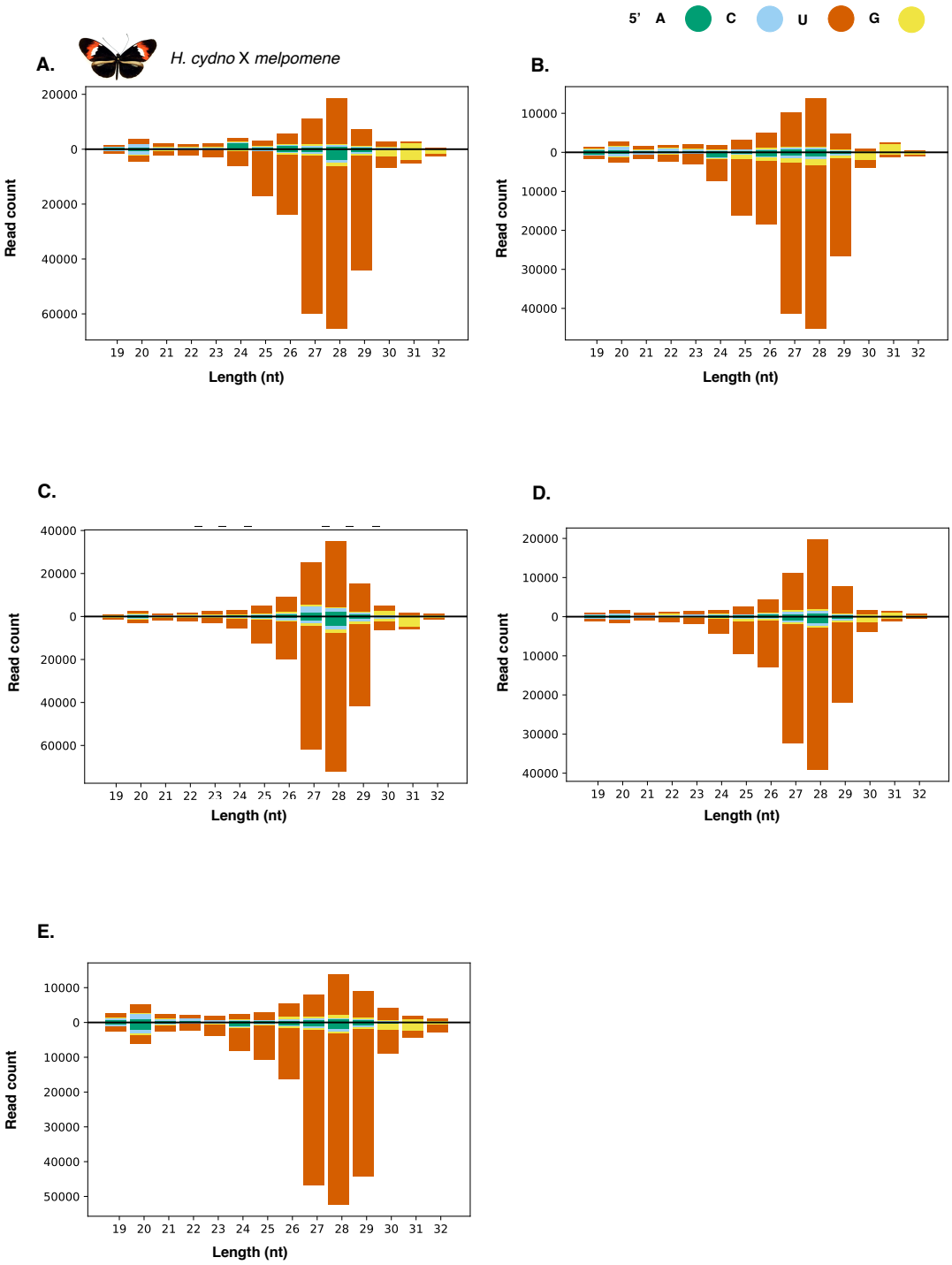

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S13. sRNAs mapping to RC TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to RC *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.

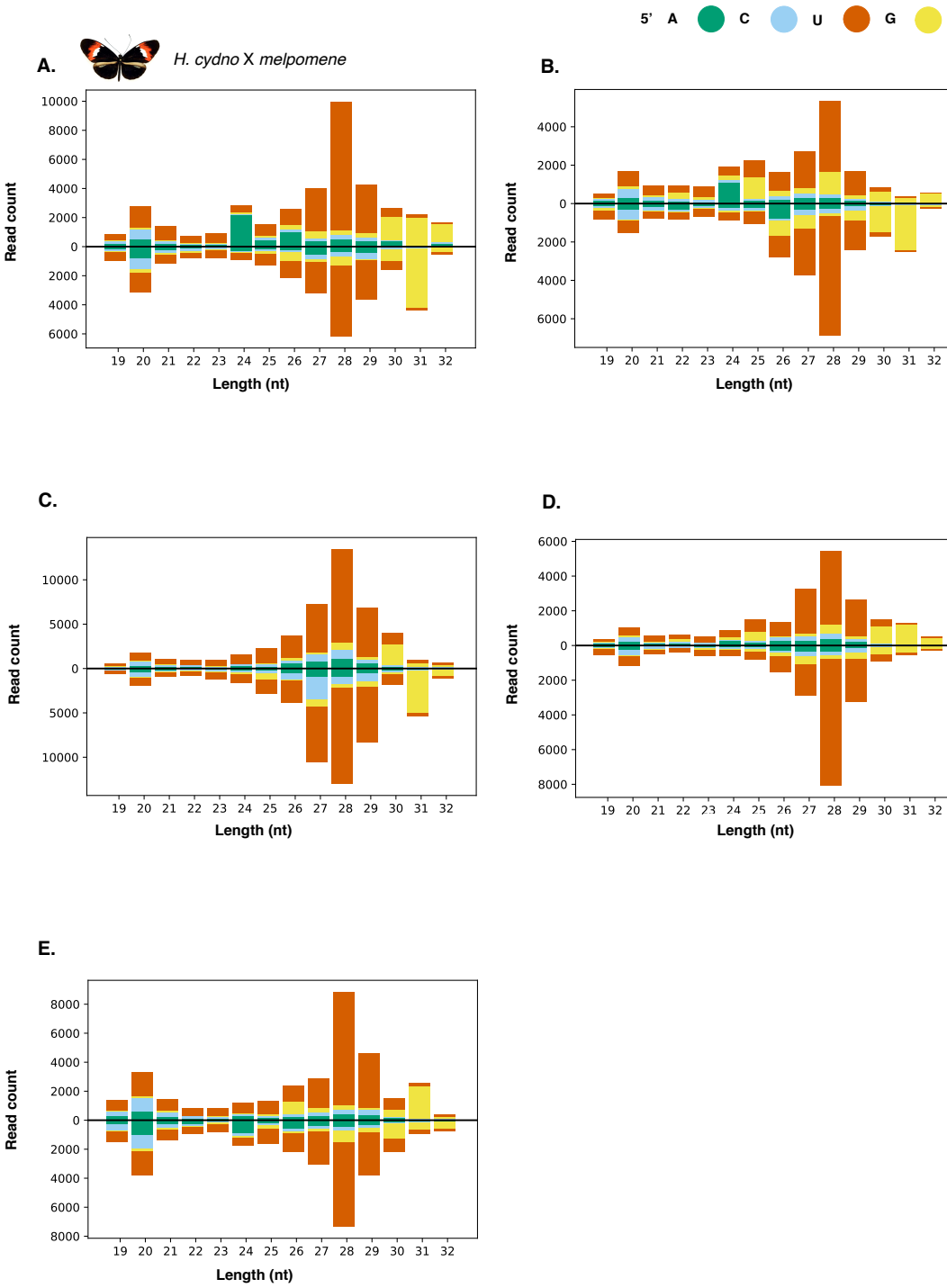
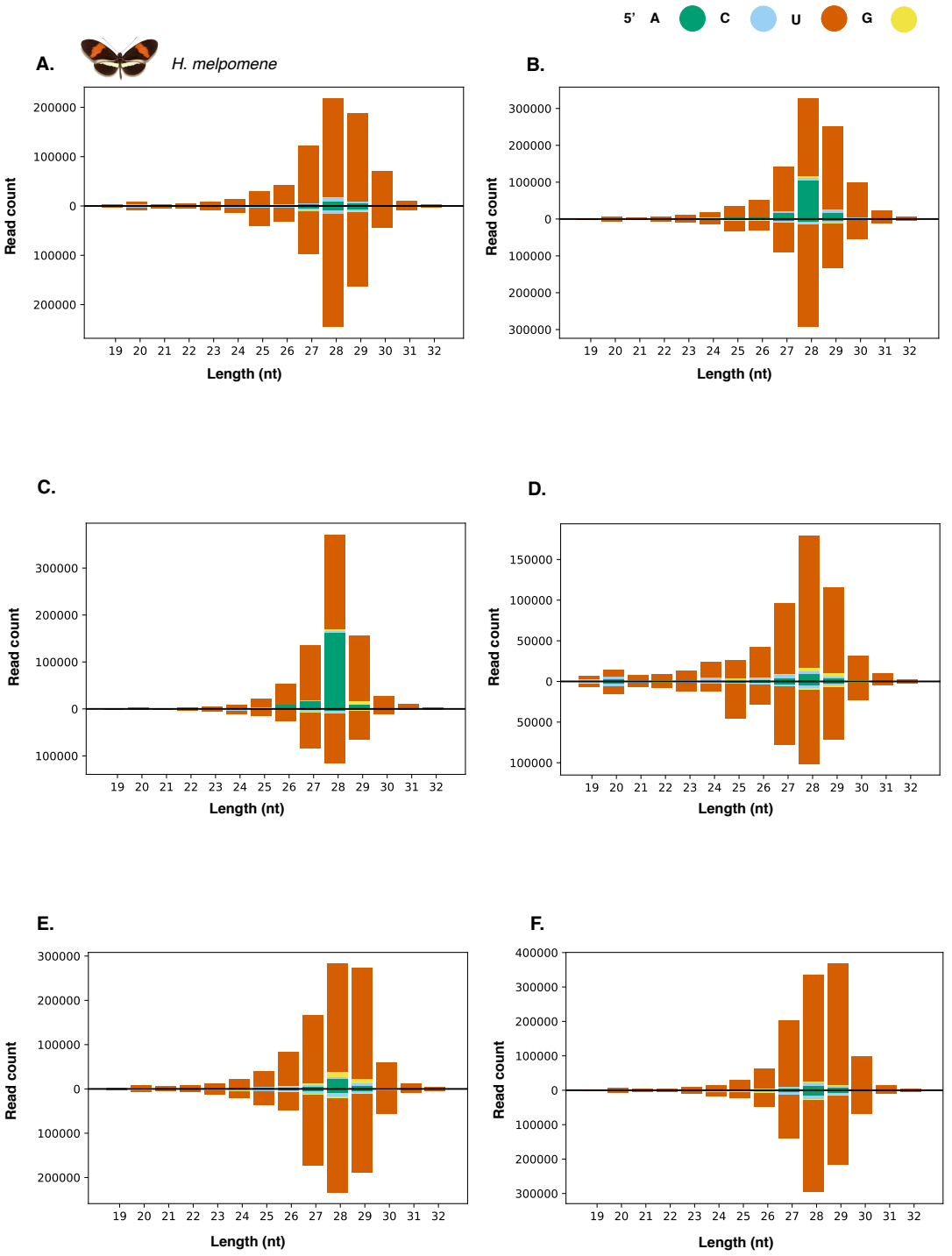

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S14. sRNAs mapping to RC TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to RC *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.


5' A
C
U
G
A.
H. melpomene
B.
C.
D.
E.
F.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

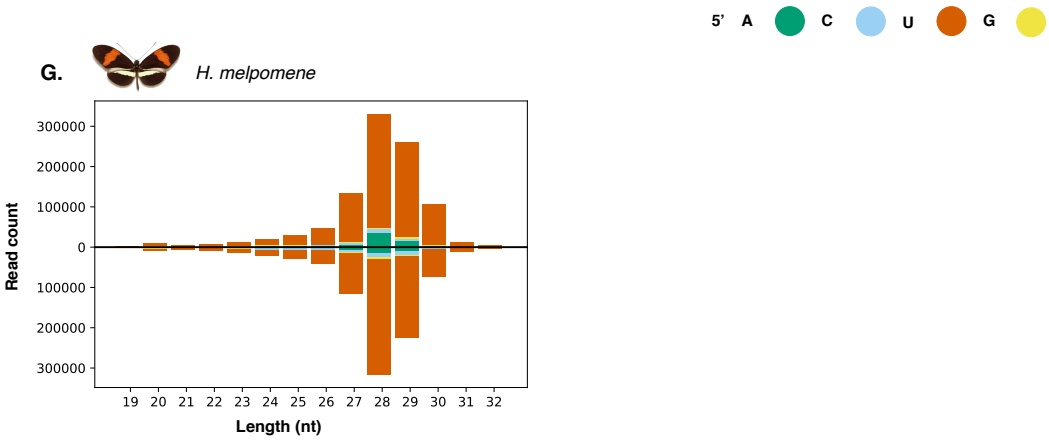


Figure S15. sRNAs mapping to LTR TEs for *H. melpomene*

sRNA read distribution for *H. melpomene* samples mapped to LTR *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33.

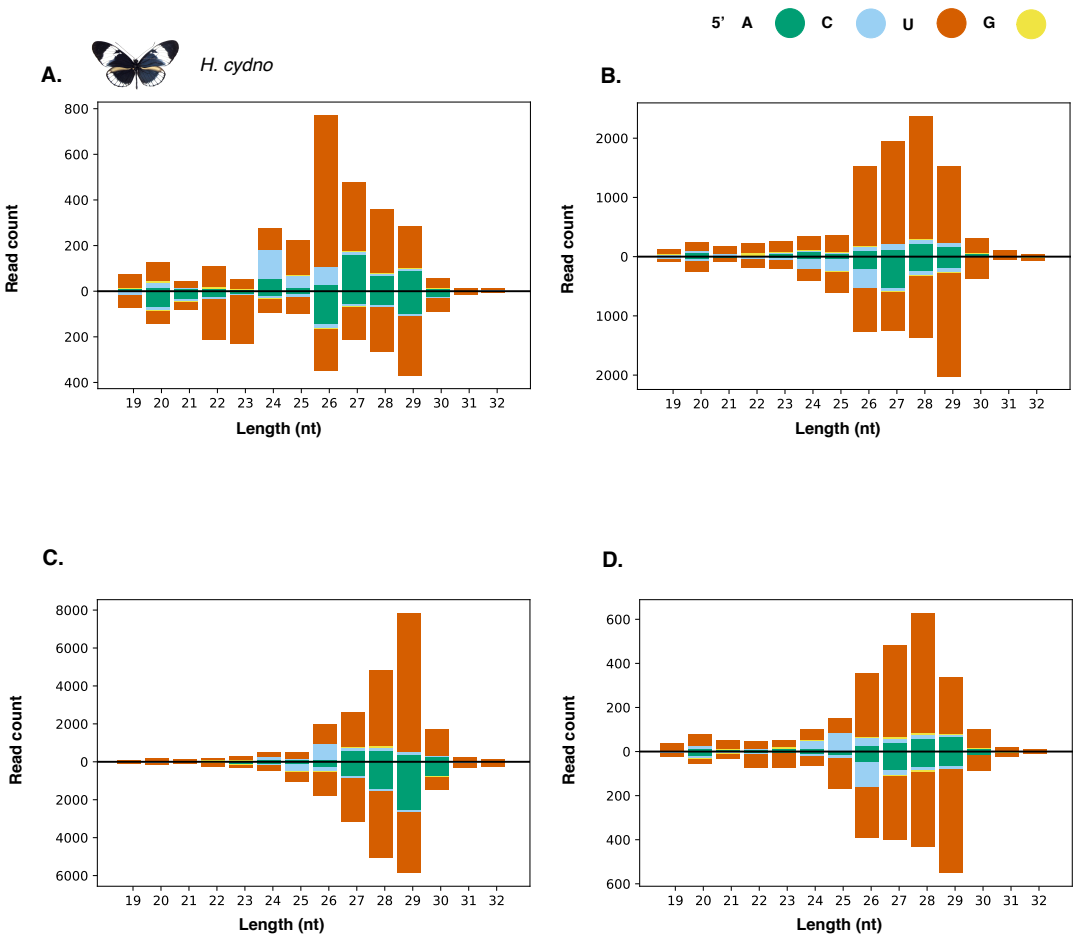


Figure S16. sRNAs mapping to LTR TEs for *H. cydno*

sRNA read distribution for *H. cydno* samples mapped to LTR *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP5, **B.** Sample AP6, **C.** Sample AP10, **D.** Sample AP17.

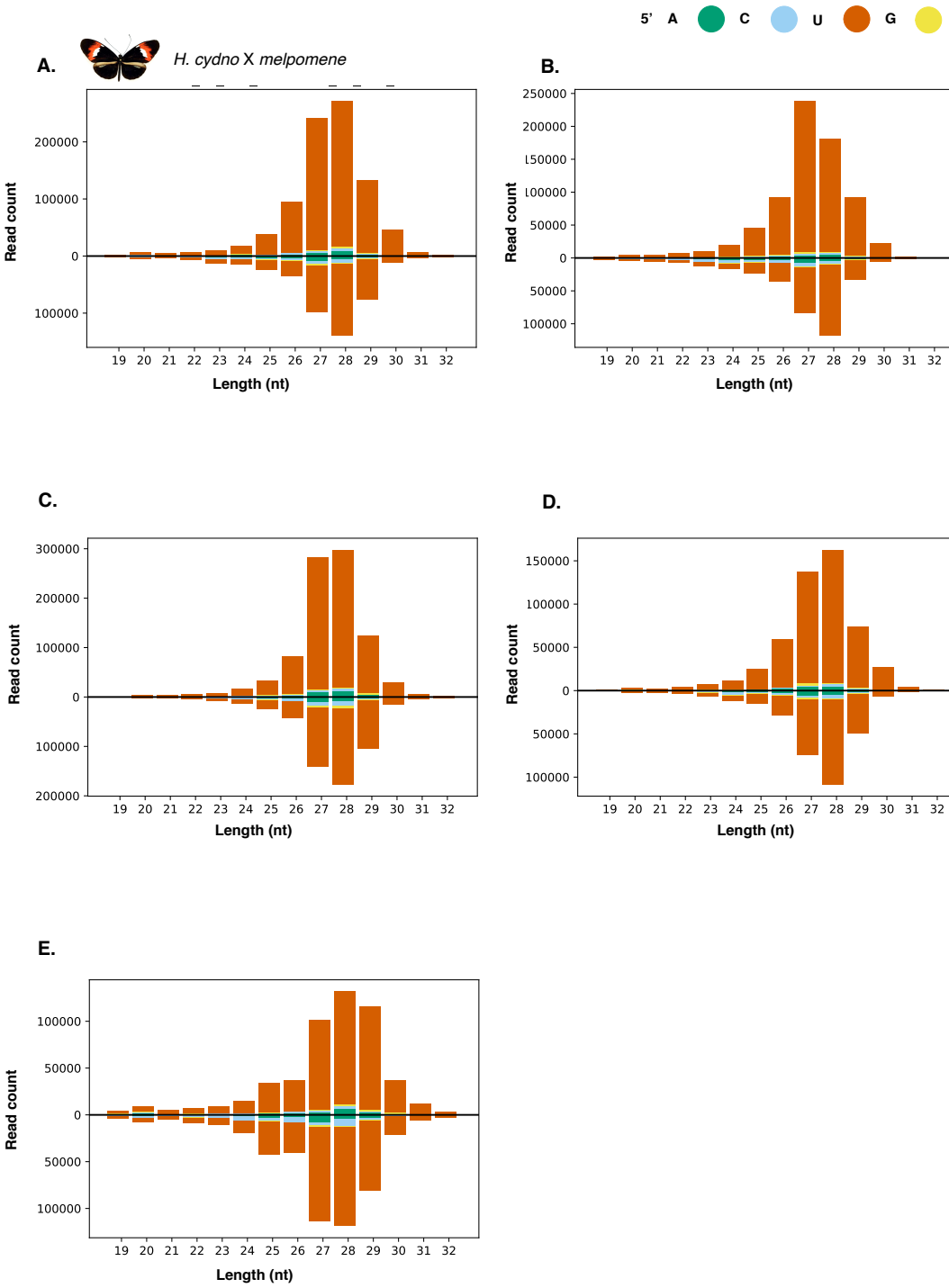

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S17. sRNAs mapping to LTR TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to LTR *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.

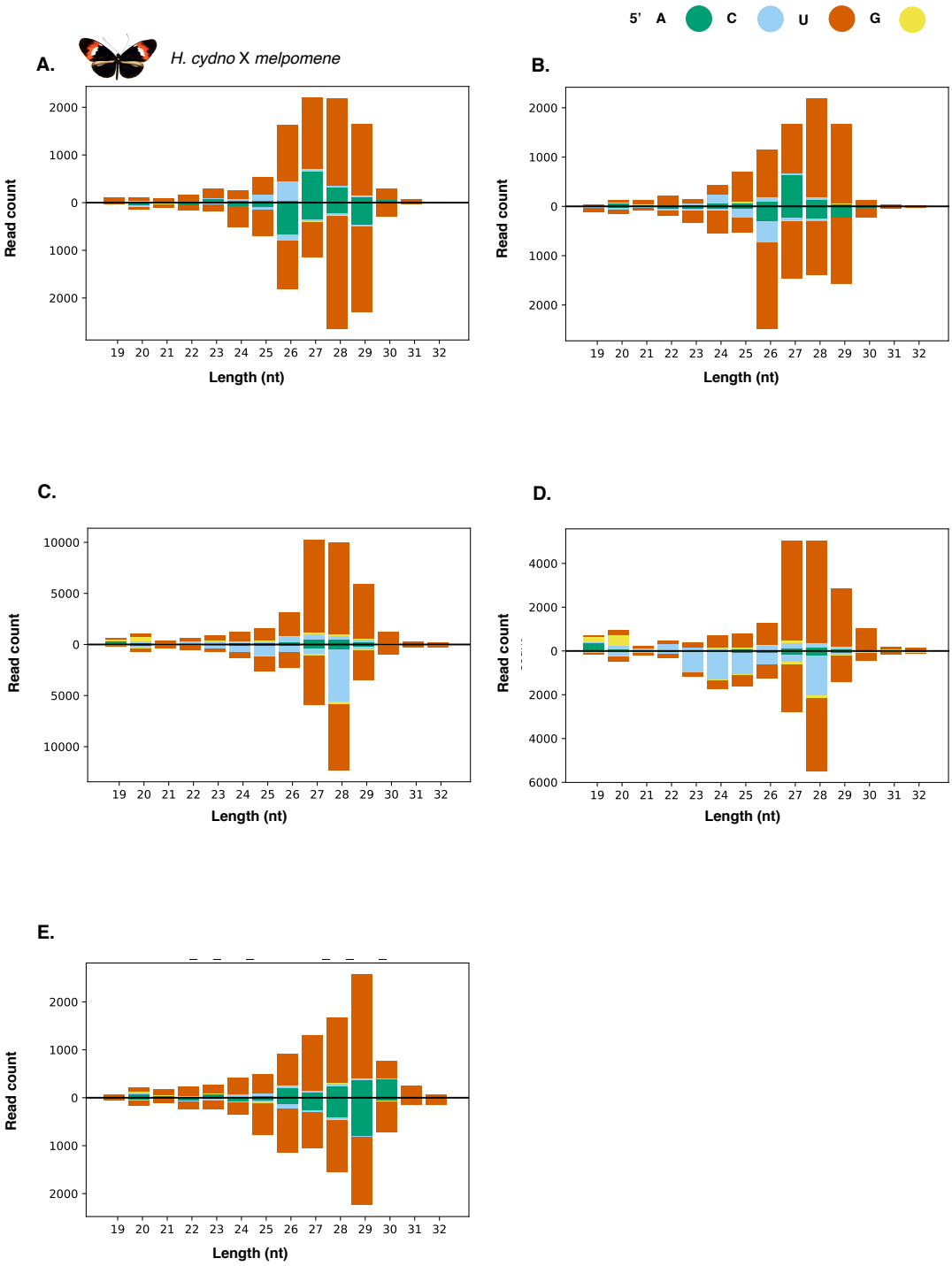

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S18. sRNAs mapping to LTR TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to LTR *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.

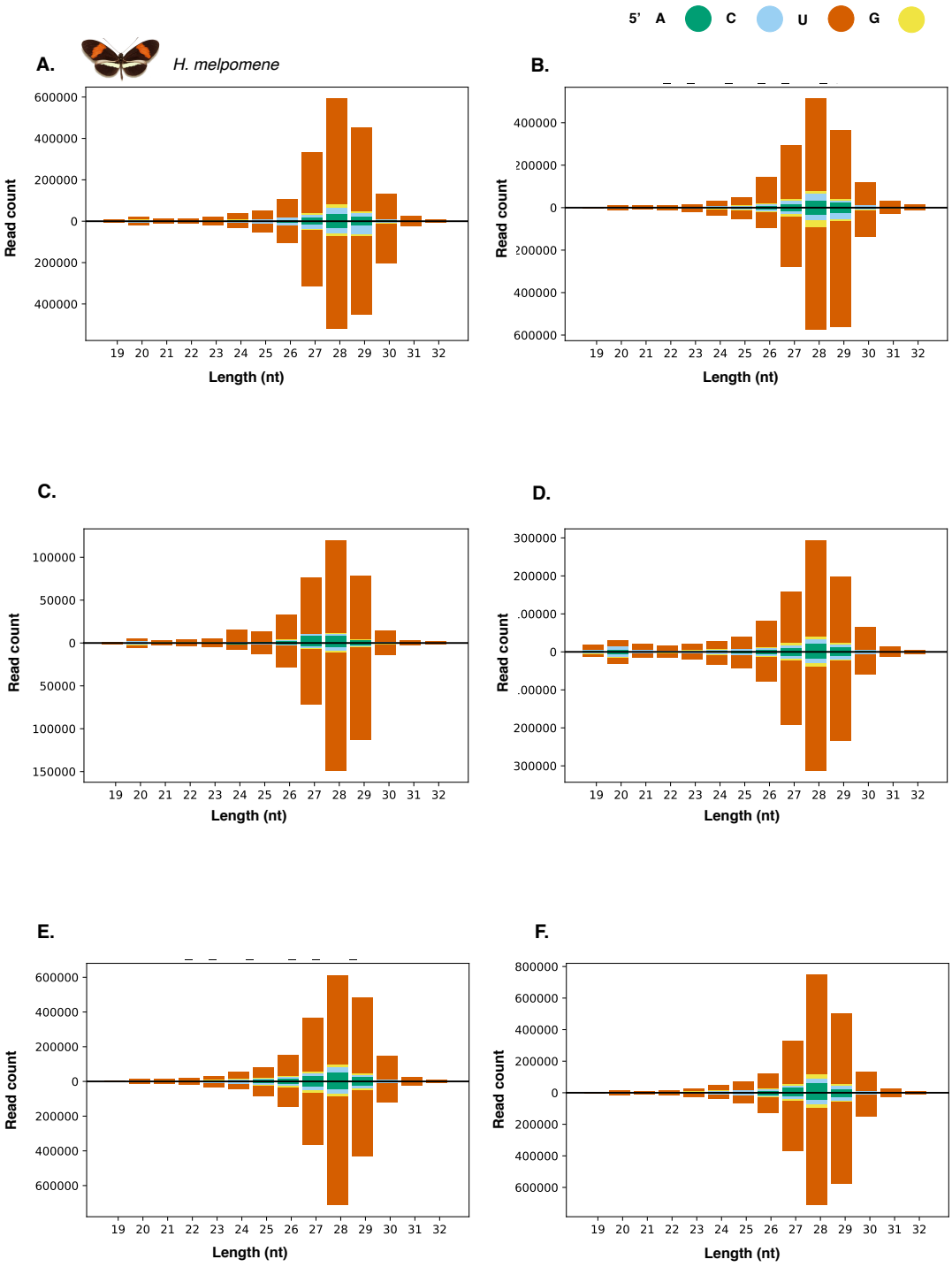

5' A C U G
A.
H. melpomene
B.
C.
D.
E.
F.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S19. sRNAs mapping to LINE TEs for *H. melpomene*

sRNA read distribution for *H. melpomene* samples mapped to LINE *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33.

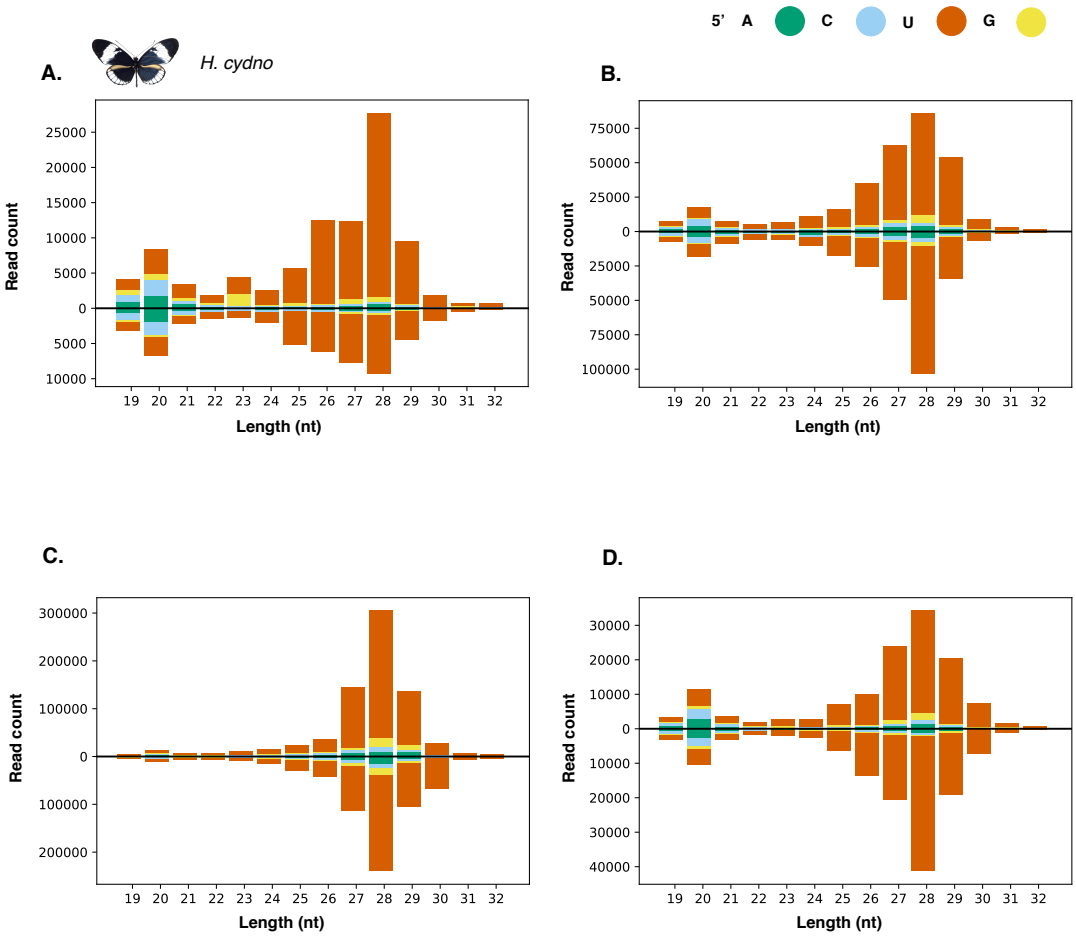


Figure S20. sRNAs mapping to LINE TEs for *H. cydno*

sRNA read distribution for *H. cydno* samples mapped to LINE *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP5, **B.** Sample AP6, **C.** Sample AP10, **D.** Sample AP17.

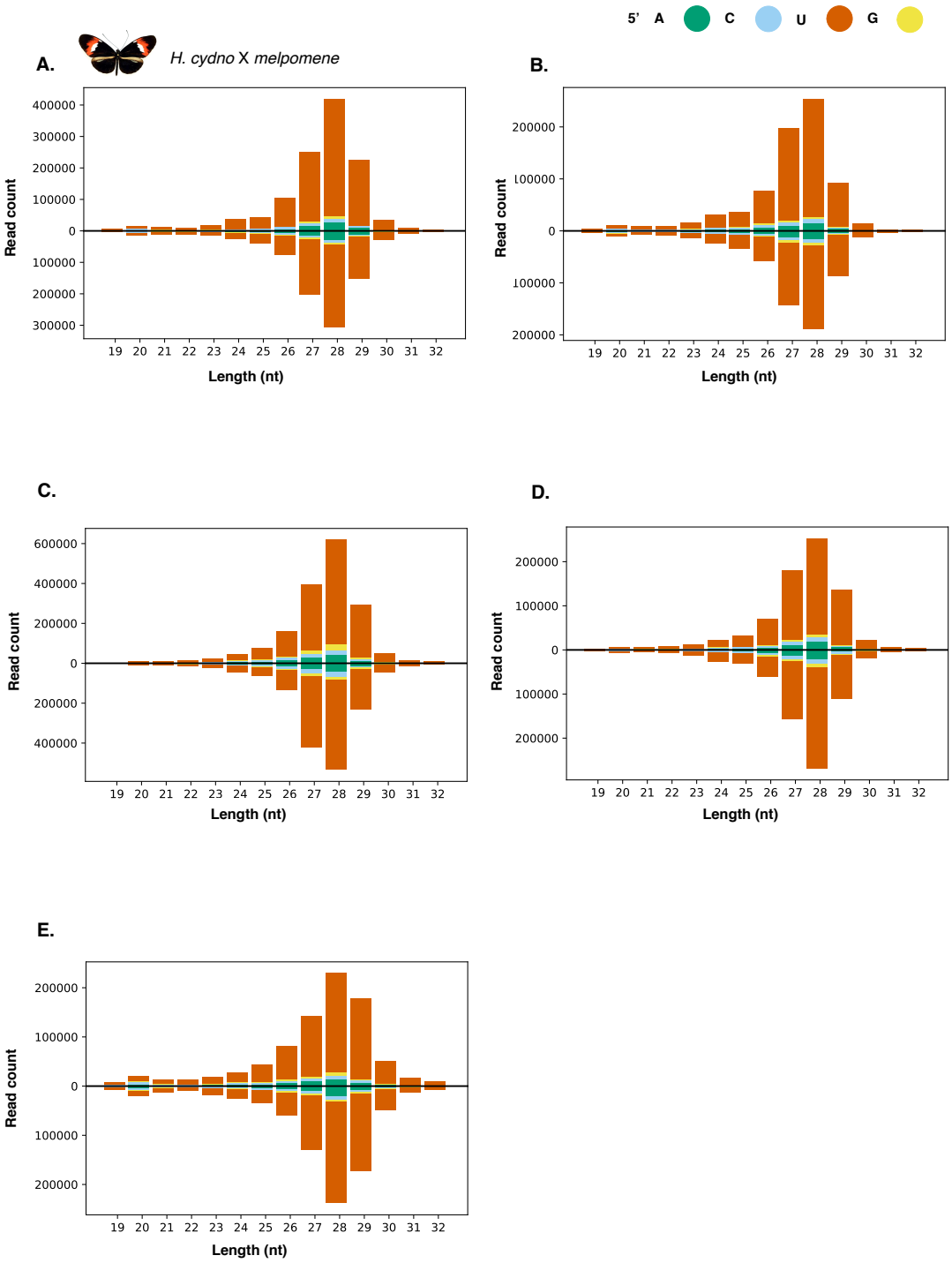

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S21. sRNAs mapping to LINE TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to LINE *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.

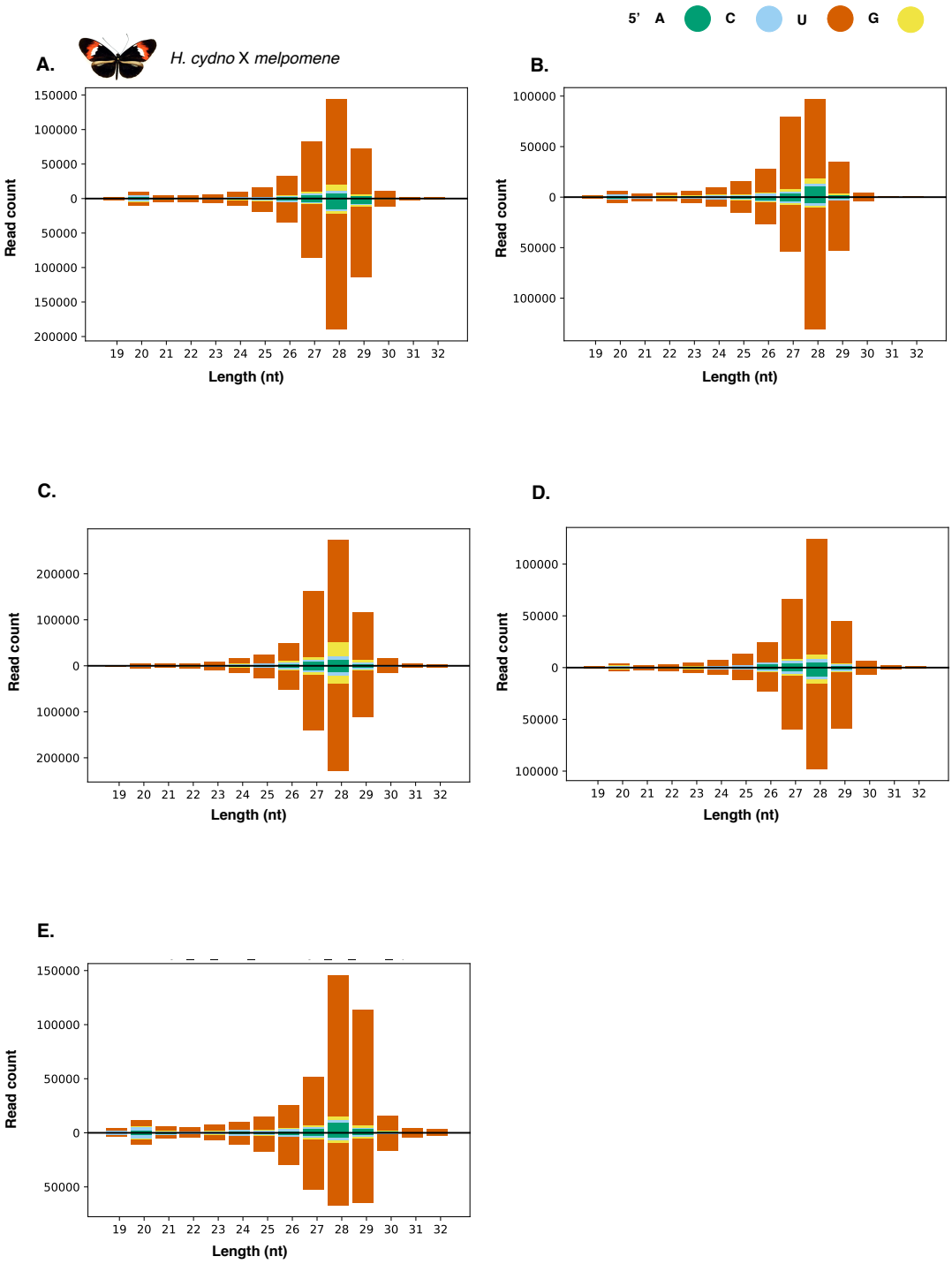
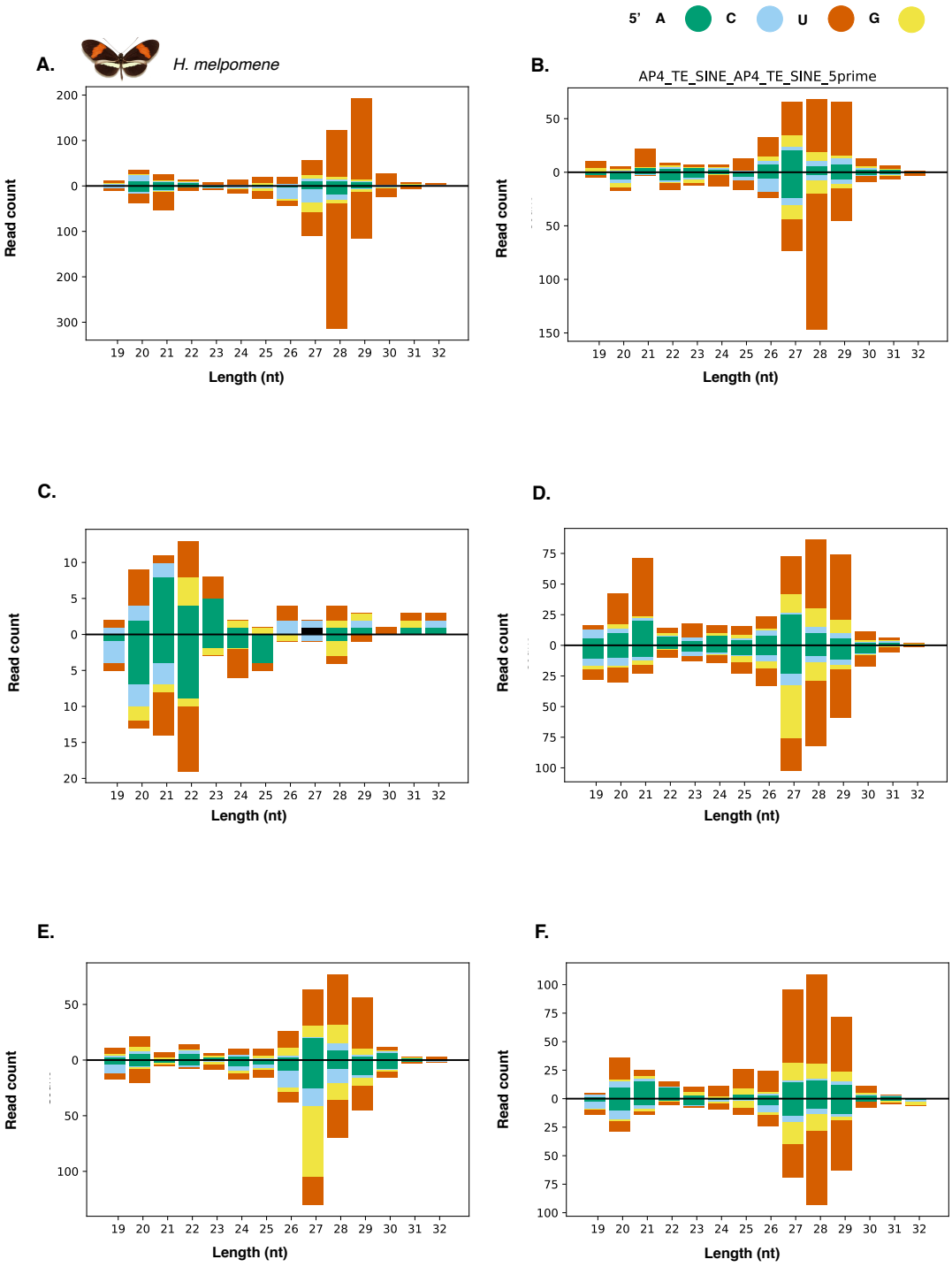

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S22. sRNAs mapping to LINE TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to LINE *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.


5' A C U G
A.
H. melpomene
B.
AP4_TE_SINE_AP4_TE_SINE_5prime
C.
D.
E.
F.
Read count
Length (nt)

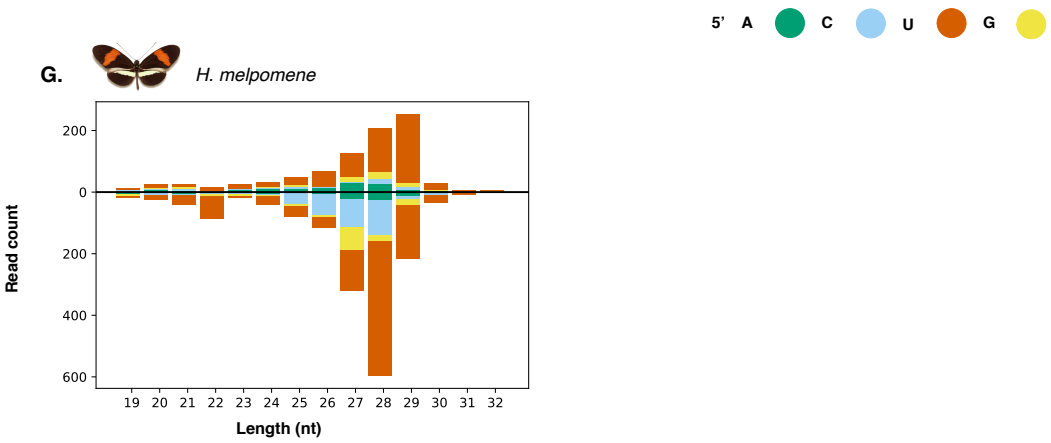


Figure S23. sRNAs mapping to SINE TEs for *H. melpomene*

sRNA read distribution for *H. melpomene* samples mapped to SINE *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33.

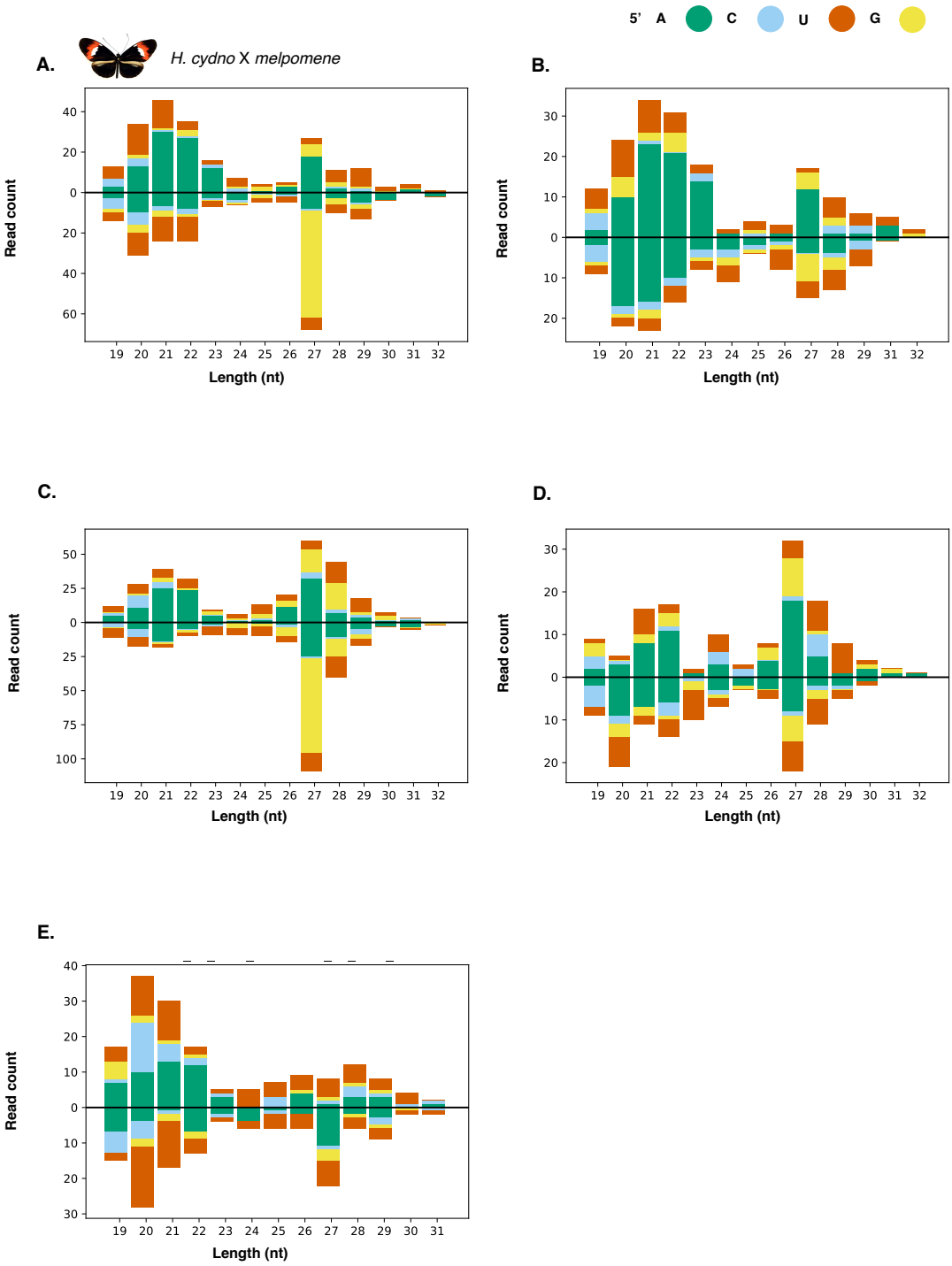
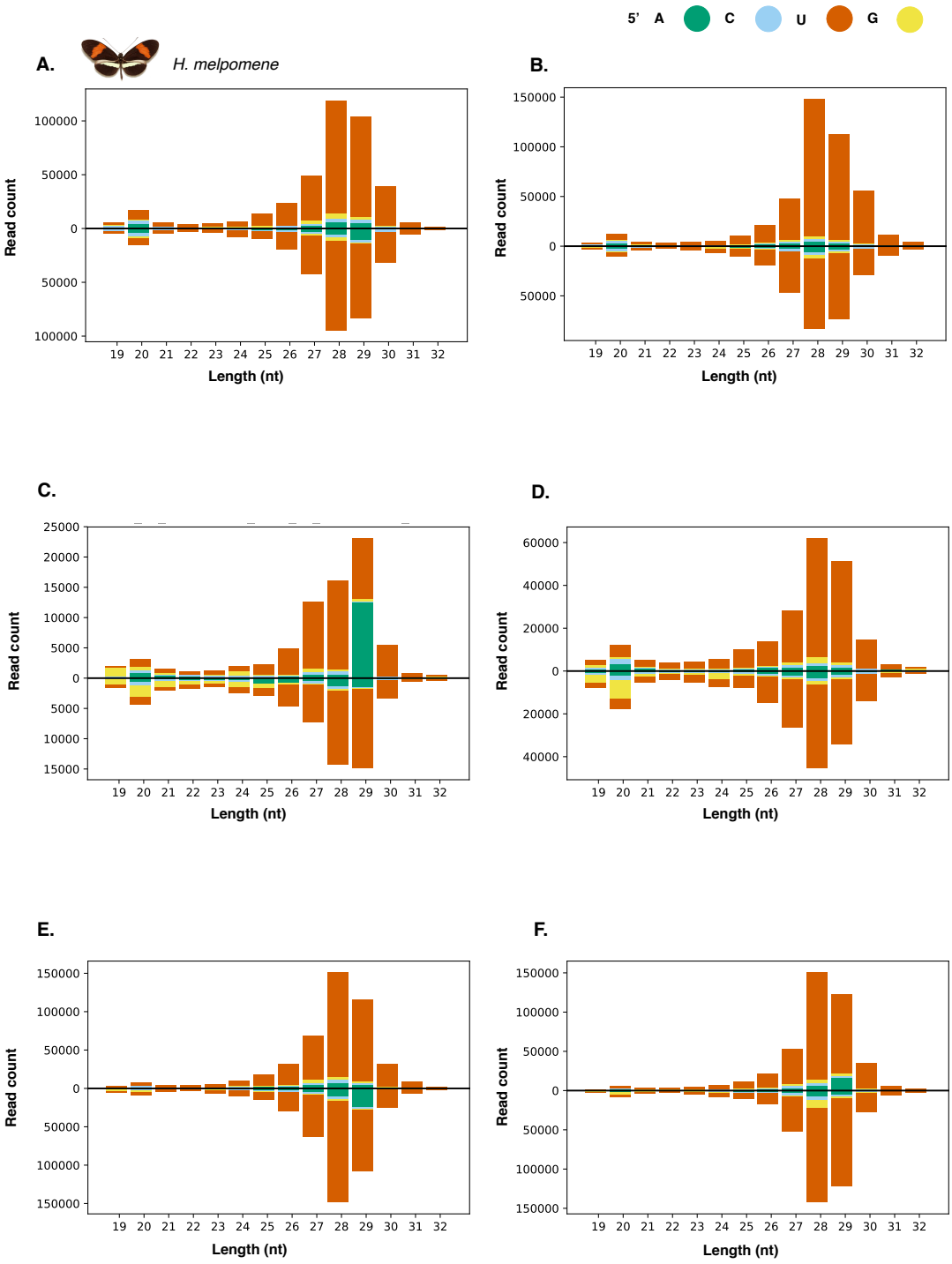

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)

Figure S24. sRNAs mapping to SINE TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to SINE *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.


5' A C U G
A.
H. melpomene
B.
C.
D.
E.
F.
Read count
Length (nt)

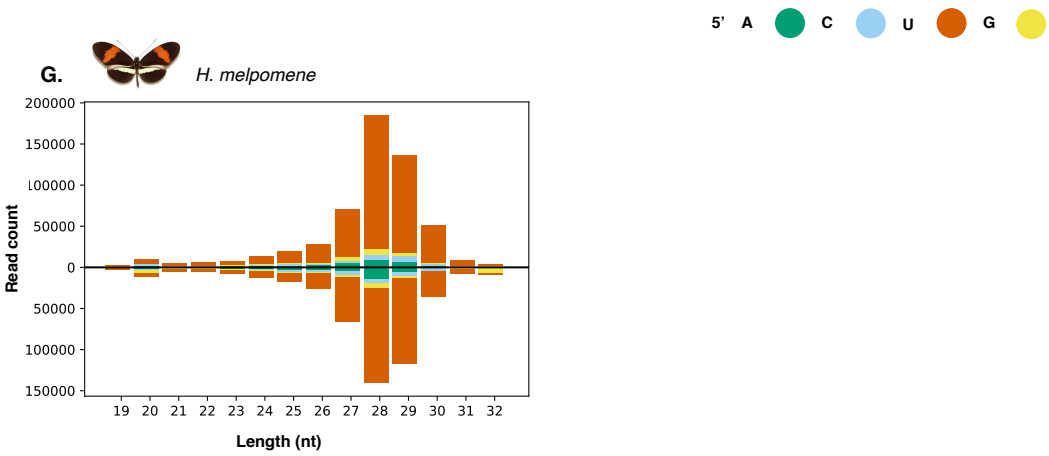


Figure S25. sRNAs mapping to unclassified TEs for *H. melpomene*.

sRNA read distribution for *H. melpomene* samples mapped to unclassified *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33.

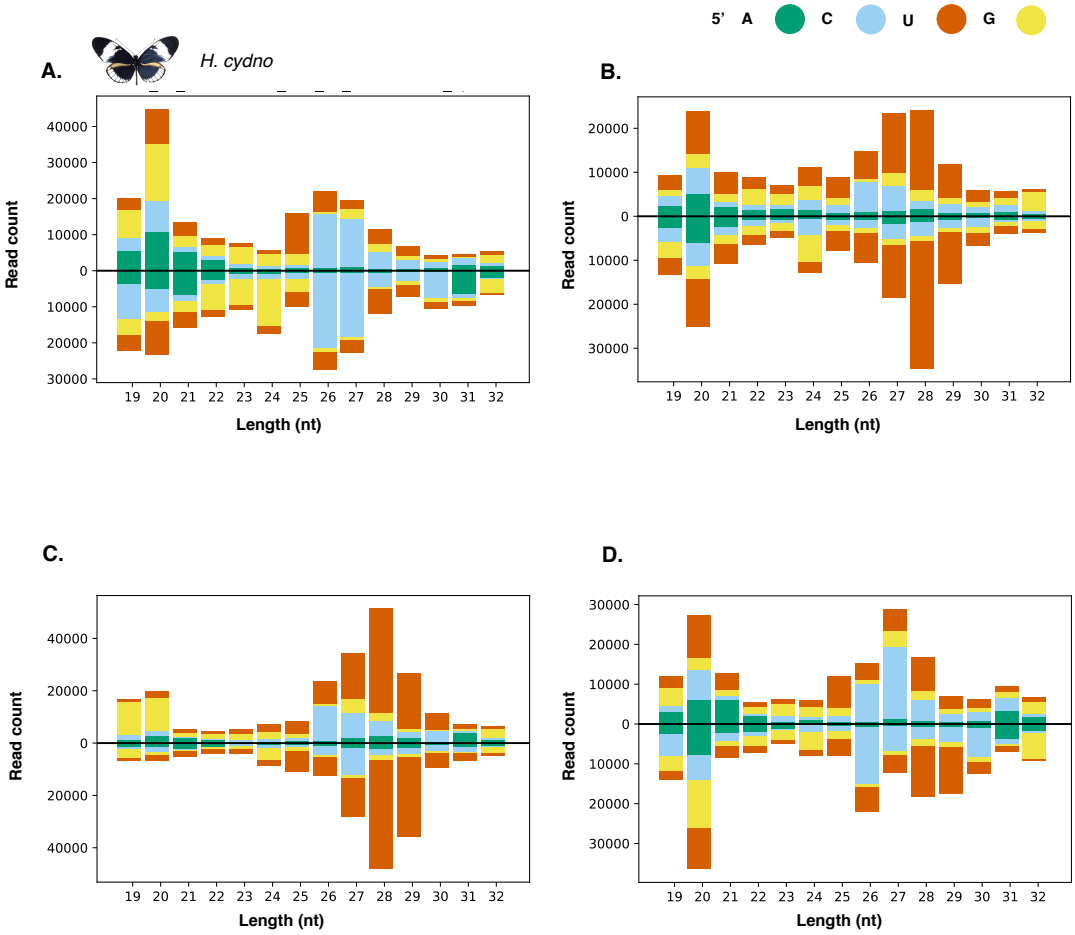


Figure S26. sRNAs mapping to unclassified TEs for *H. cydno*.

sRNA read distribution for *H. cydno* samples mapped to unclassified *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP5, **B.** Sample AP6, **C.** Sample AP10, **D.** Sample AP17.

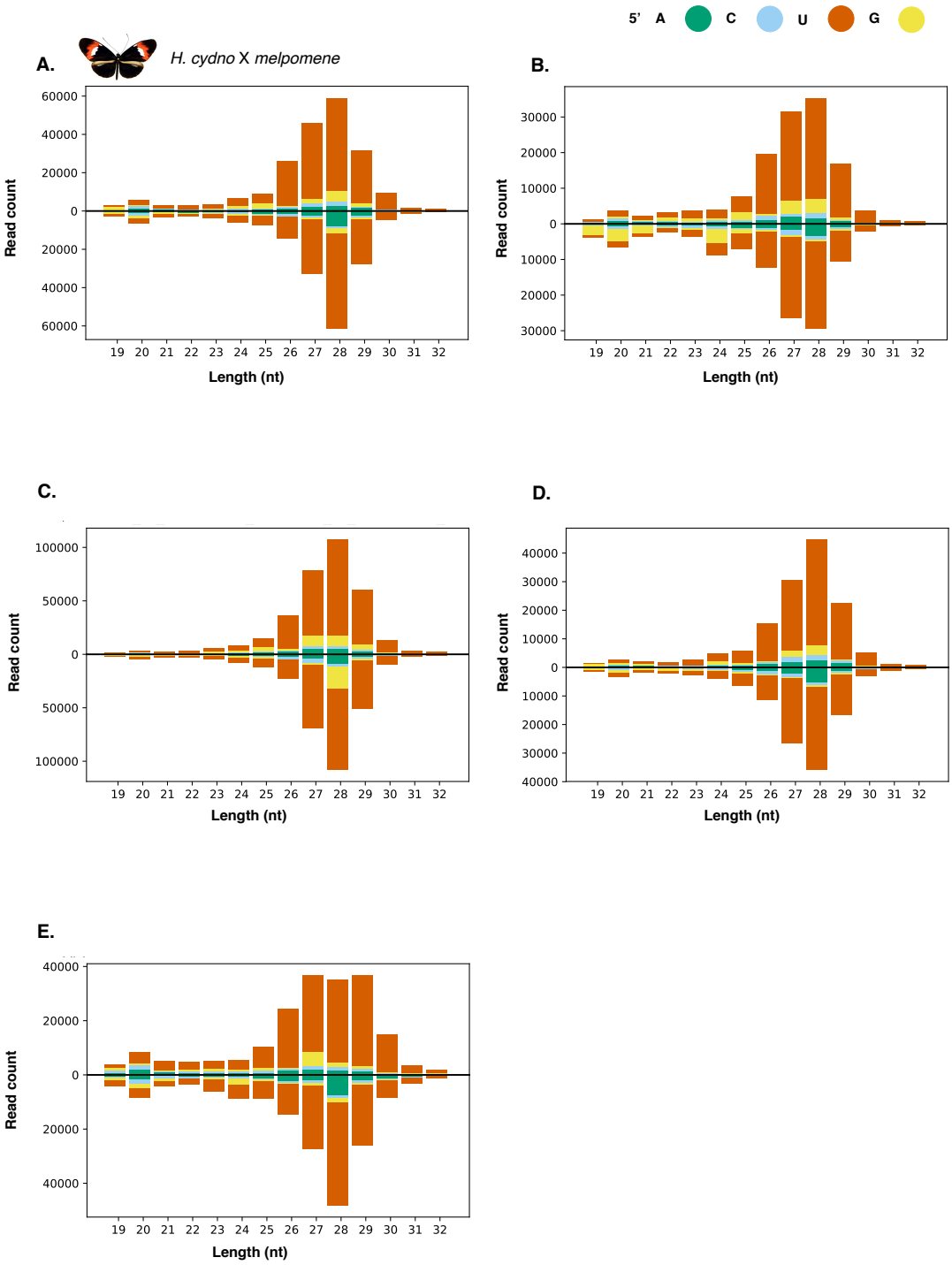

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S27. sRNAs mapping to unclassified TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to unclassified *H. melpomene* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.

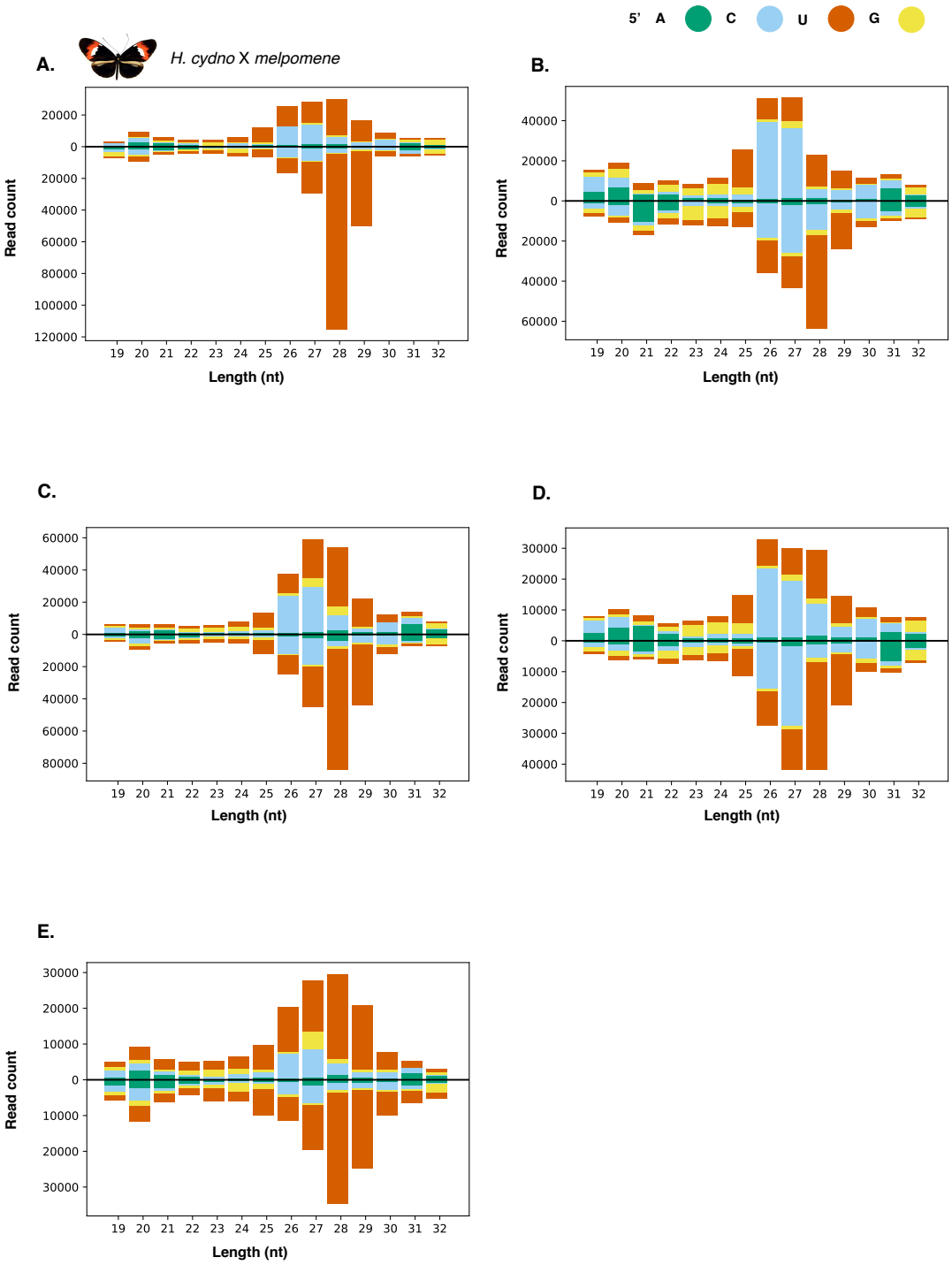
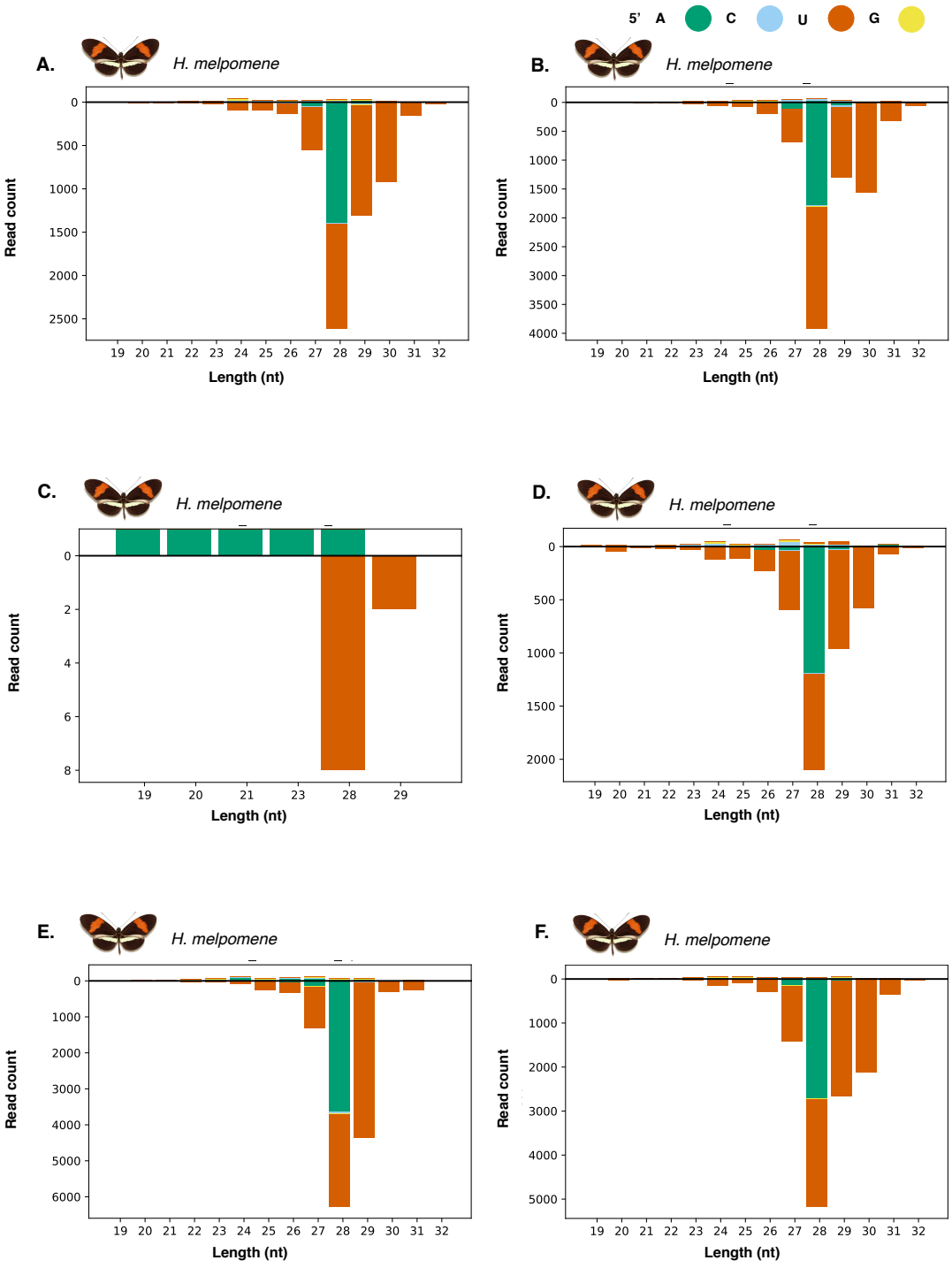

5' A C U G
A.
H. cydno X melpomene
B.
C.
D.
E.
Read count
Length (nt)
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure S28. sRNAs mapping to unclassified TEs for *H. cydno* x *melpomene* hybrids

sRNA read distribution for *H. cydno* x *melpomene* samples mapped to unclassified *H. cydno* TEs. y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. **A.** Sample AP50, **B.** Sample AP57, **C.** Sample AP59, **D.** Sample AP60, and **E.** Sample AP72.


5' A C U G
A.
H. melpomene
B.
H. melpomene
C.
H. melpomene
D.
H. melpomene
E.
H. melpomene
F.
H. melpomene
Read count
Length (nt)

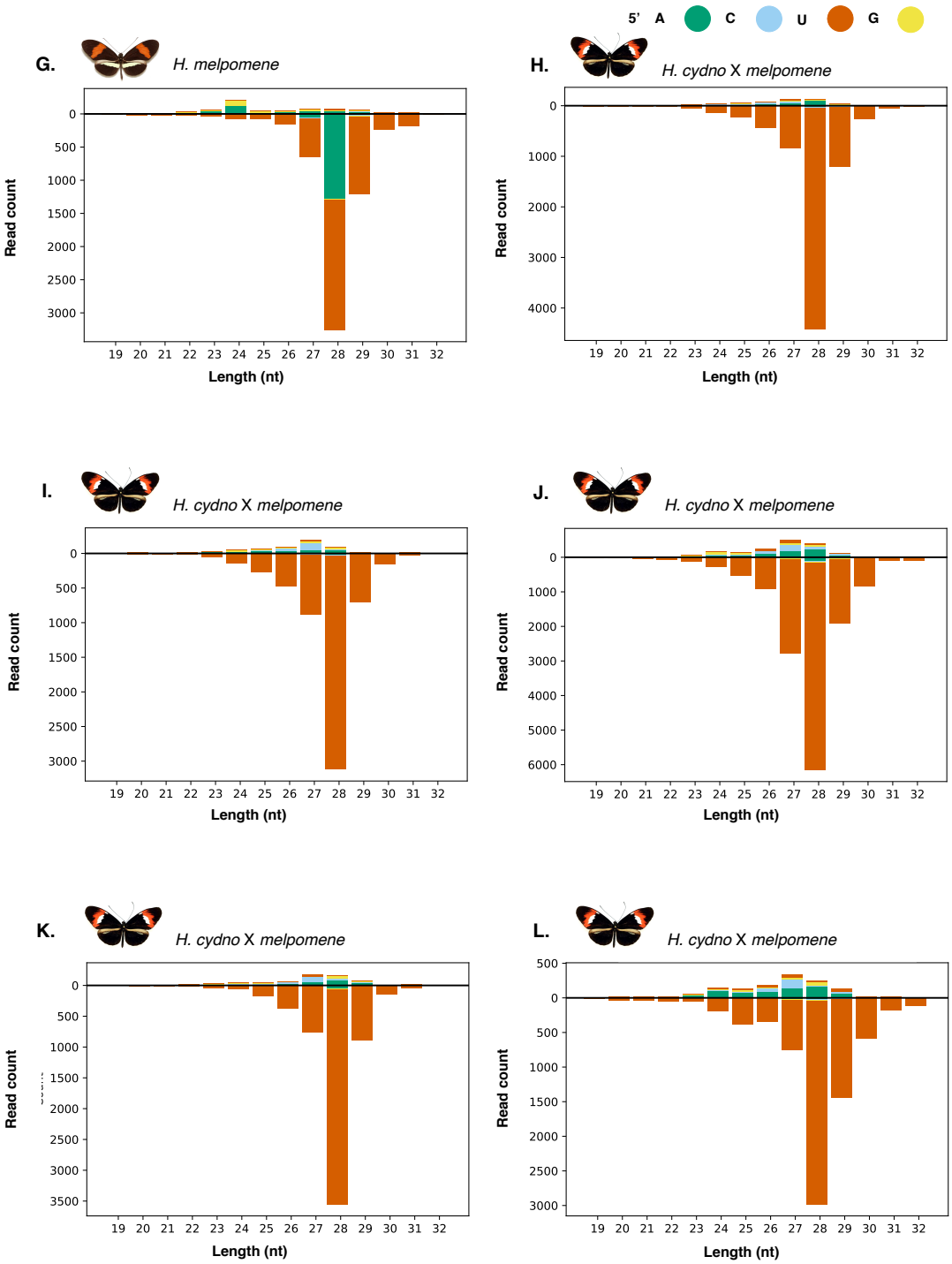
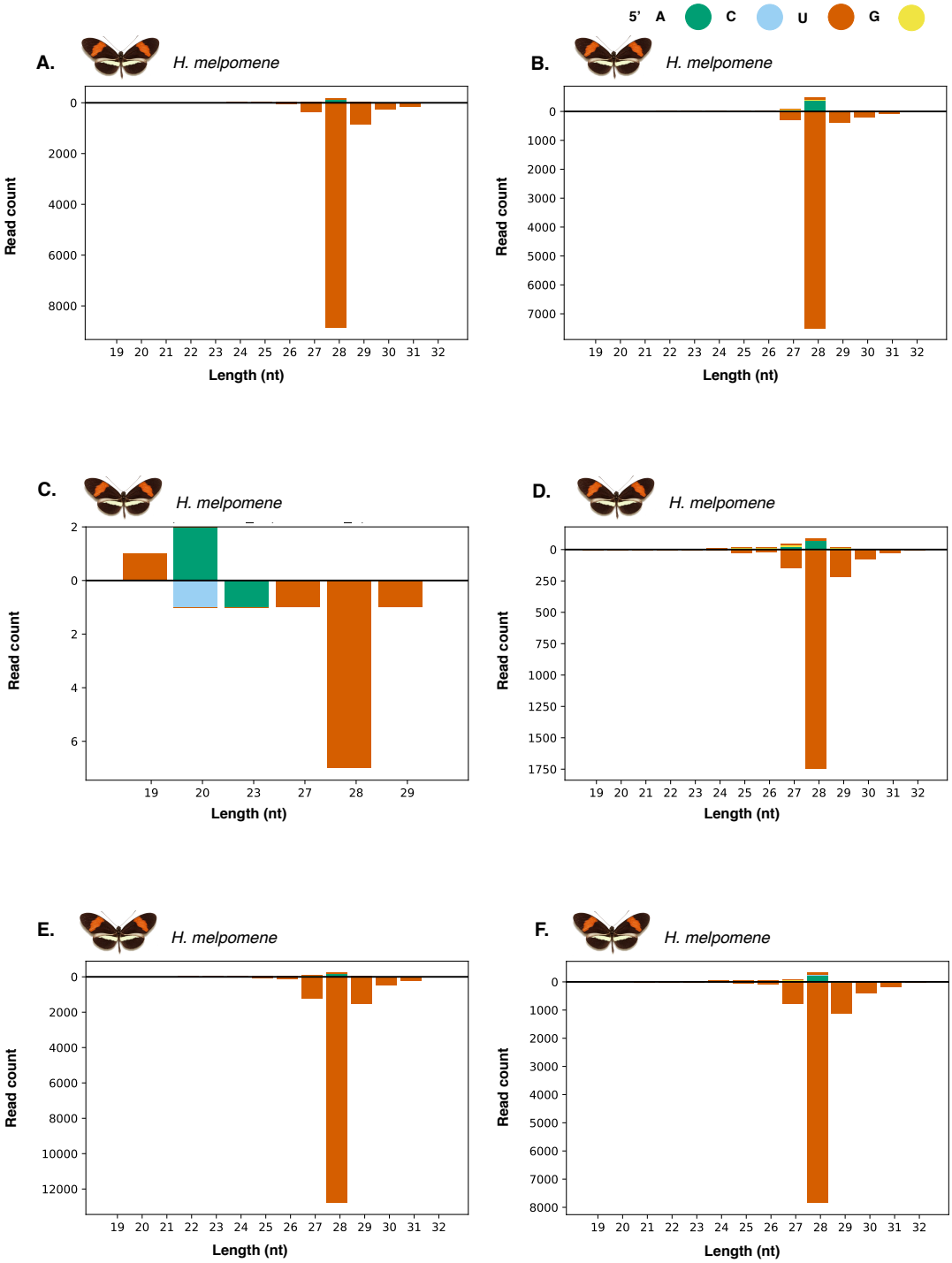

5' A C U G
G.
H. melpomene
H.
H. cydno X melpomene
I.
H. cydno X melpomene
J.
H. cydno X melpomene
K.
H. cydno X melpomene
L.
H. cydno X melpomene
Read count
Length (nt)

Figure S29. *H. melpomene* and hybrid sRNAs mapping to *H. melpomene*'s BEL-2 TE

sRNA read distribution for *H. melpomene* and *H. cydno* x *H. melpomene* samples mapped to *H. melpomene*'s BEL-2 TE. *H. melpomene*'s BEL-2 TE is significantly more abundant in hybrid samples than *H. melpomene* samples (Figure 4, Table 2). y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. *H. melpomene* samples: **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33. *H. cydno* x *H. melpomene* samples: **H.** Sample AP50, **I.** Sample AP57, **J.** Sample AP59, **K.** Sample AP60, and **L.** Sample AP72.


5' A C U G
A.
H. melpomene
B.
H. melpomene
C.
H. melpomene
D.
H. melpomene
E.
H. melpomene
F.
H. melpomene
Read count
Length (nt)

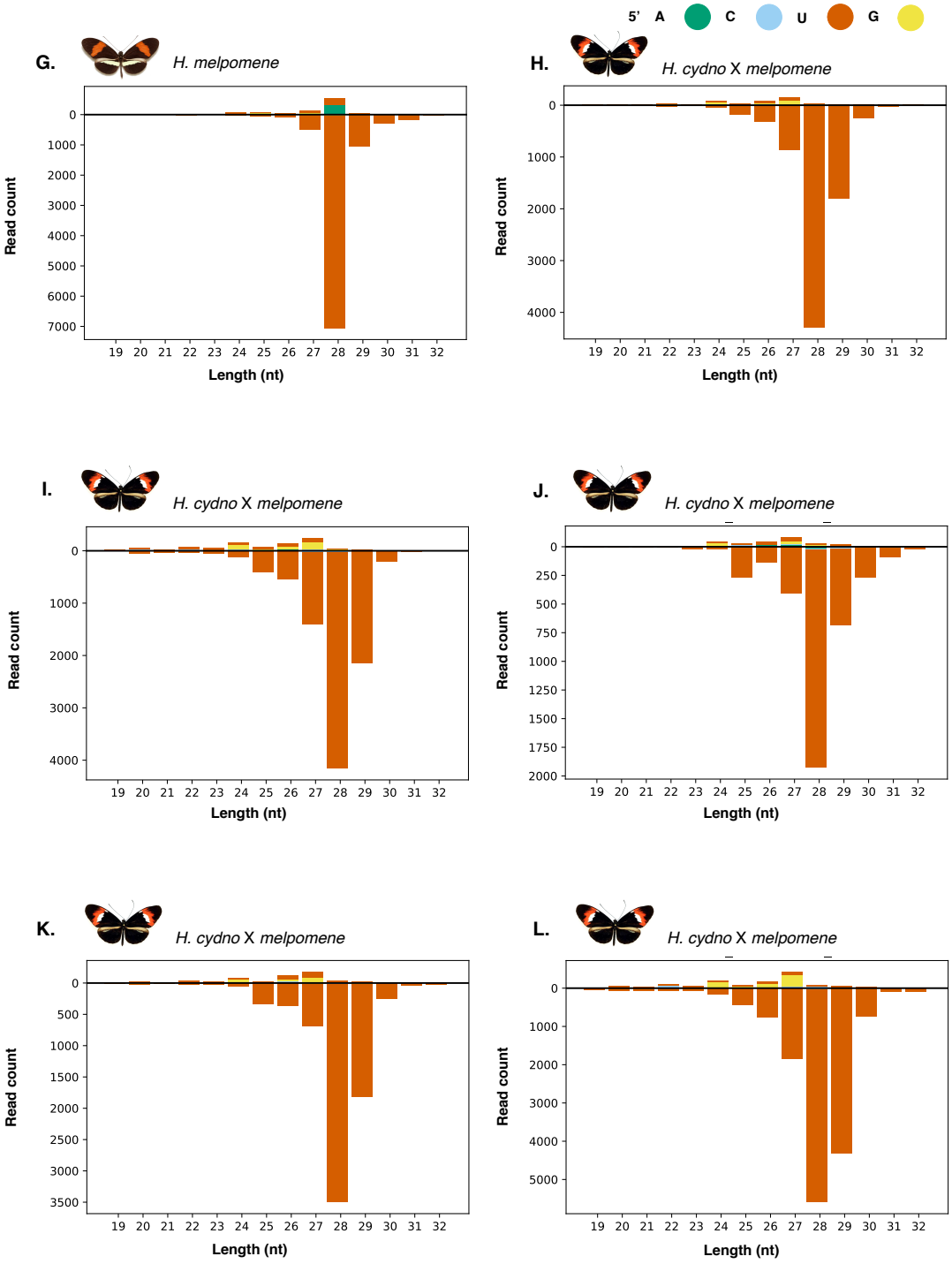
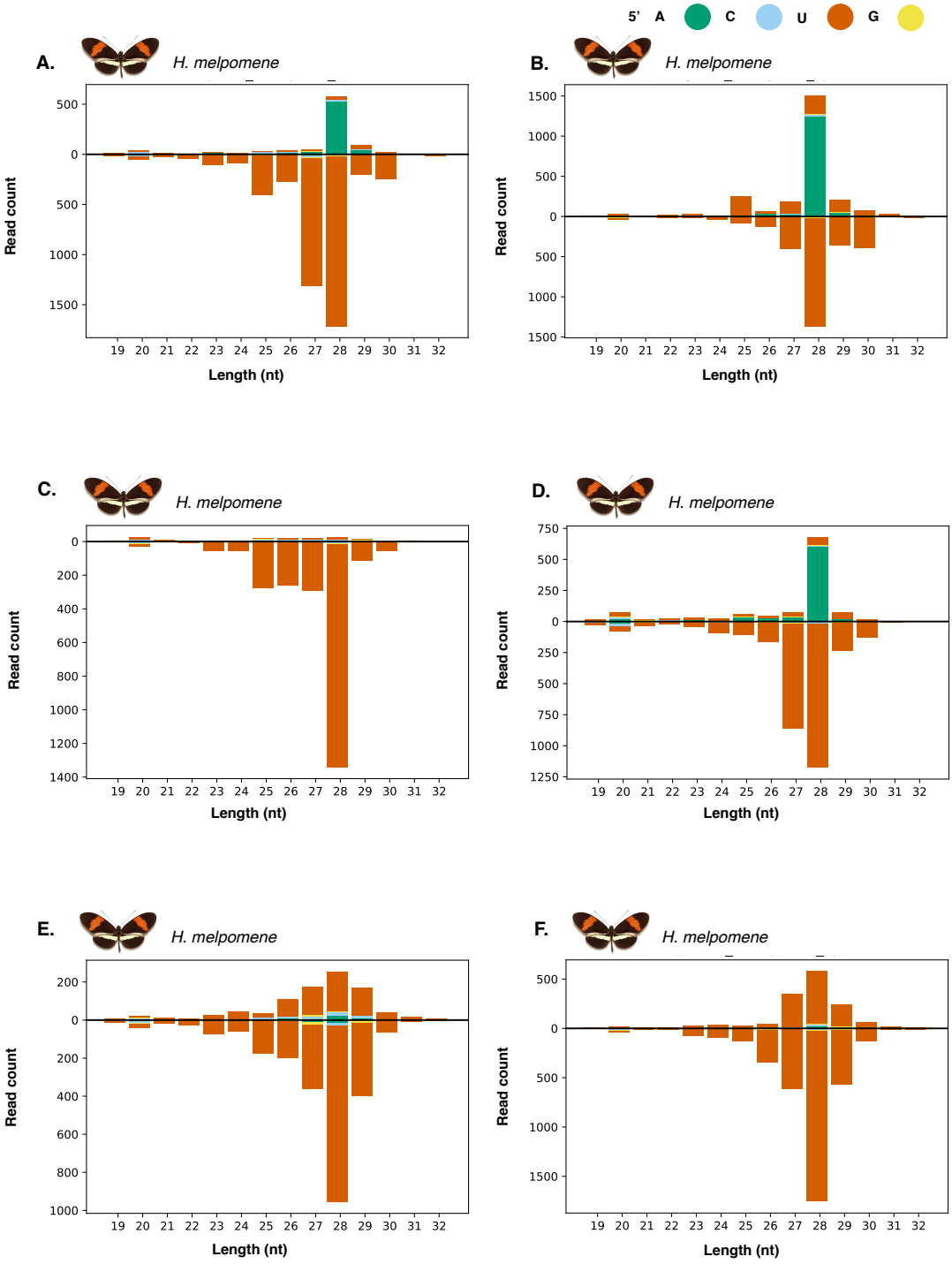

5' A C U G
G.
H. melpomene
H.
H. cydno X melpomene
I.
H. cydno X melpomene
J.
H. cydno X melpomene
K.
H. cydno X melpomene
L.
H. cydno X melpomene
Read count
Length (nt)

Figure S30. *H. melpomene* and hybrid sRNAs mapping to *H. melpomene*'s Copia-6 TE

sRNA read distribution for *H. melpomene* and *H. cydno* x *H. melpomene* samples mapped to *H. melpomene*'s Copia-6 TE. *H. melpomene*'s Copia-6 TE is significantly more abundant in hybrid samples than *H. melpomene* samples (Figure 4, Table 2). y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. *H. melpomene* samples: **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33. *H. cydno* x *H. melpomene* samples: **H.** Sample AP50, **I.** Sample AP57, **J.** Sample AP59, **K.** Sample AP60, and **L.** Sample AP72.


5' A C U G
A.
H. melpomene
B.
H. melpomene
C.
H. melpomene
D.
H. melpomene
E.
H. melpomene
F.
H. melpomene
Read count
Length (nt)

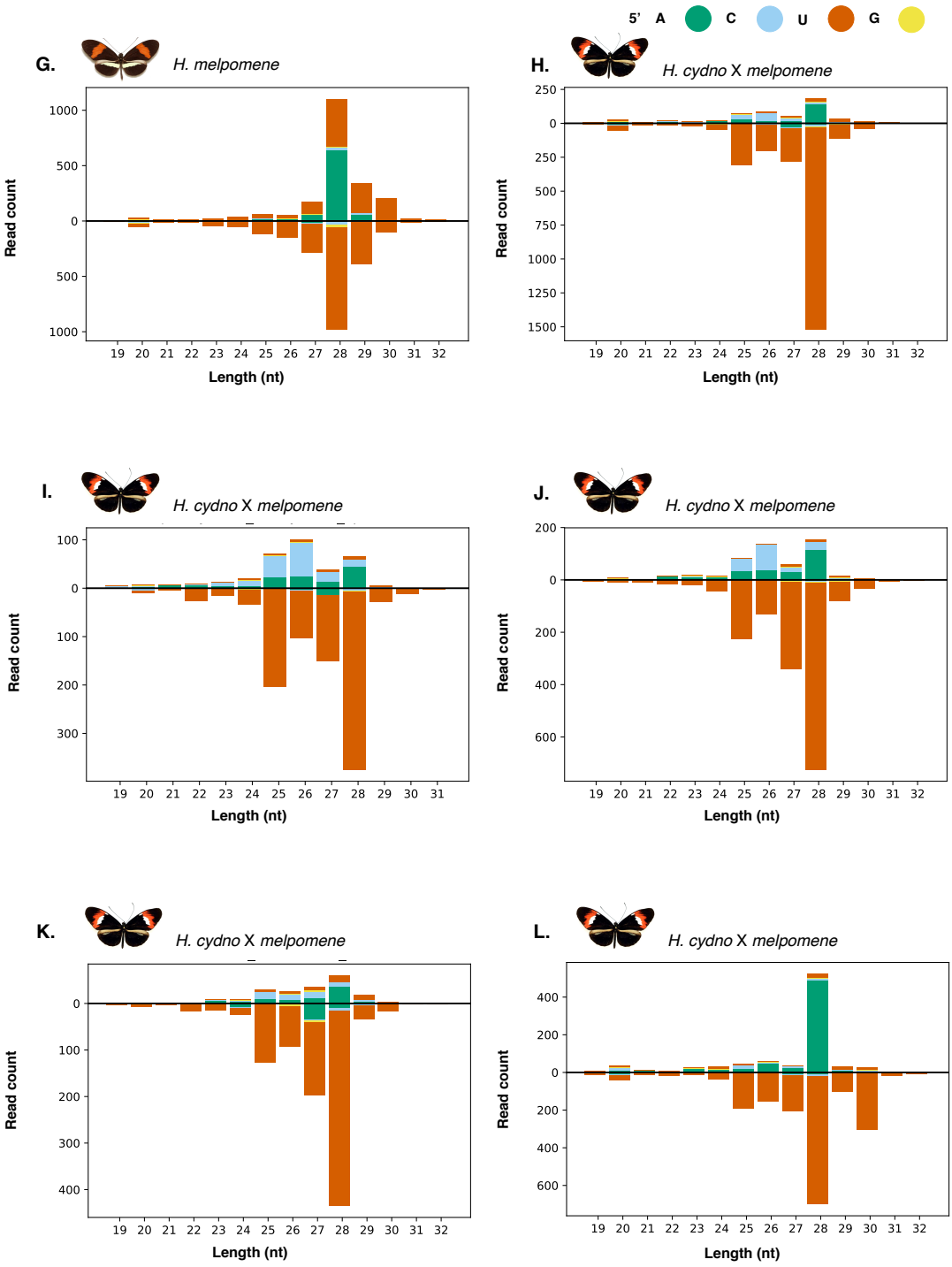
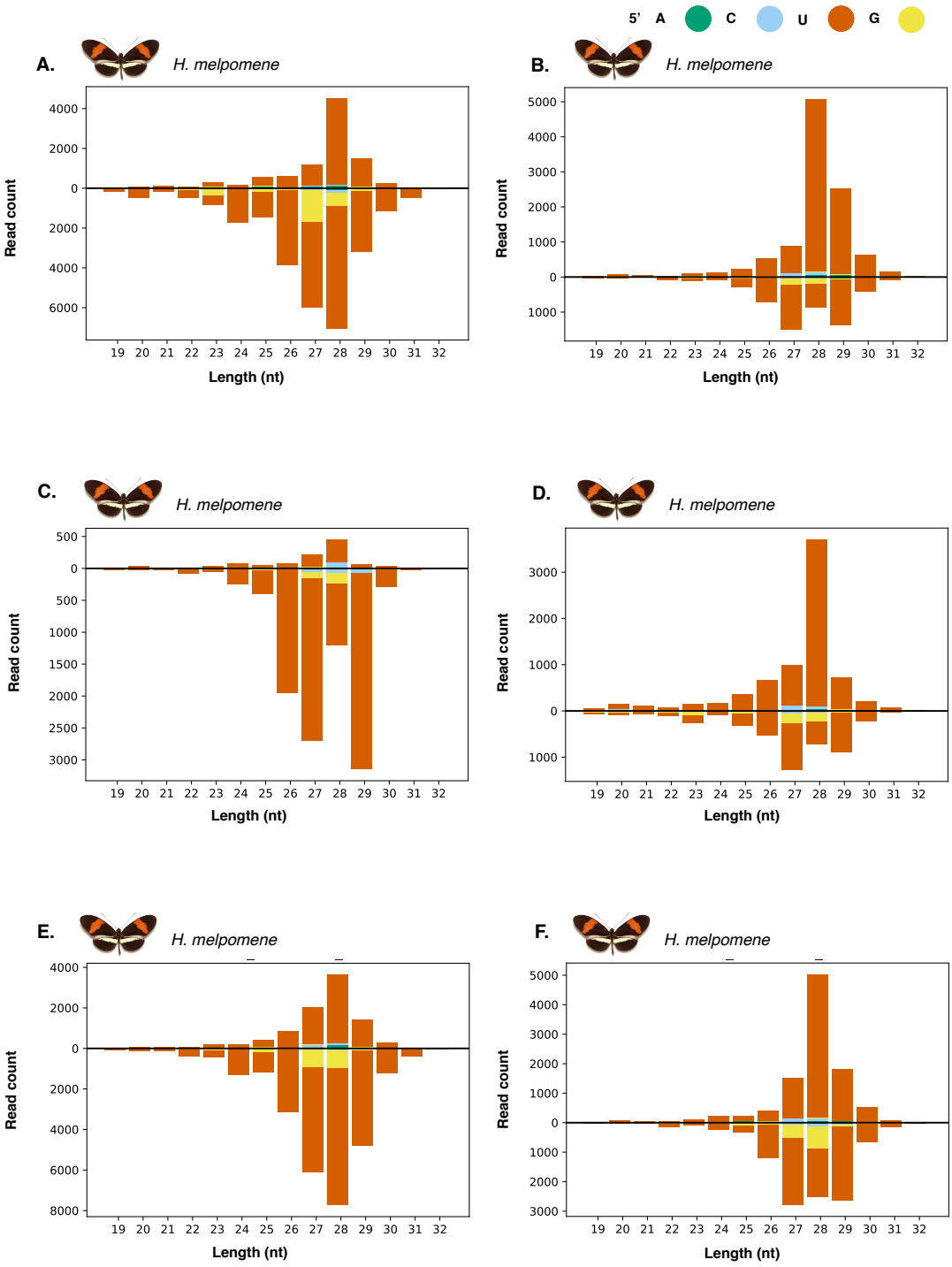

5' A C U G
G.
H. melpomene
H.
H. cydno X melpomene
Read count
Length (nt)
I.
H. cydno X melpomene
J.
H. cydno X melpomene
K.
H. cydno X melpomene
L.
H. cydno X melpomene

Figure S31. *H. melpomene* and hybrid sRNAs mapping to *H. melpomene*'s Jockey TE

sRNA read distribution for *H. melpomene* and *H. cydno* x *H. melpomene* samples mapped to *H. melpomene*'s Jockey TE. *H. melpomene*'s Jockey TE is significantly more abundant in hybrid samples than *H. melpomene* samples (Figure 4, Table 2). y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. *H. melpomene* samples: **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33. *H. cydno* x *H. melpomene* samples: **H.** Sample AP50, **I.** Sample AP57, **J.** Sample AP59, **K.** Sample AP60, and **L.** Sample AP72.


5' A C U G
A.
H. melpomene
B.
H. melpomene
C.
H. melpomene
D.
H. melpomene
E.
H. melpomene
F.
H. melpomene
Read count
Length (nt)

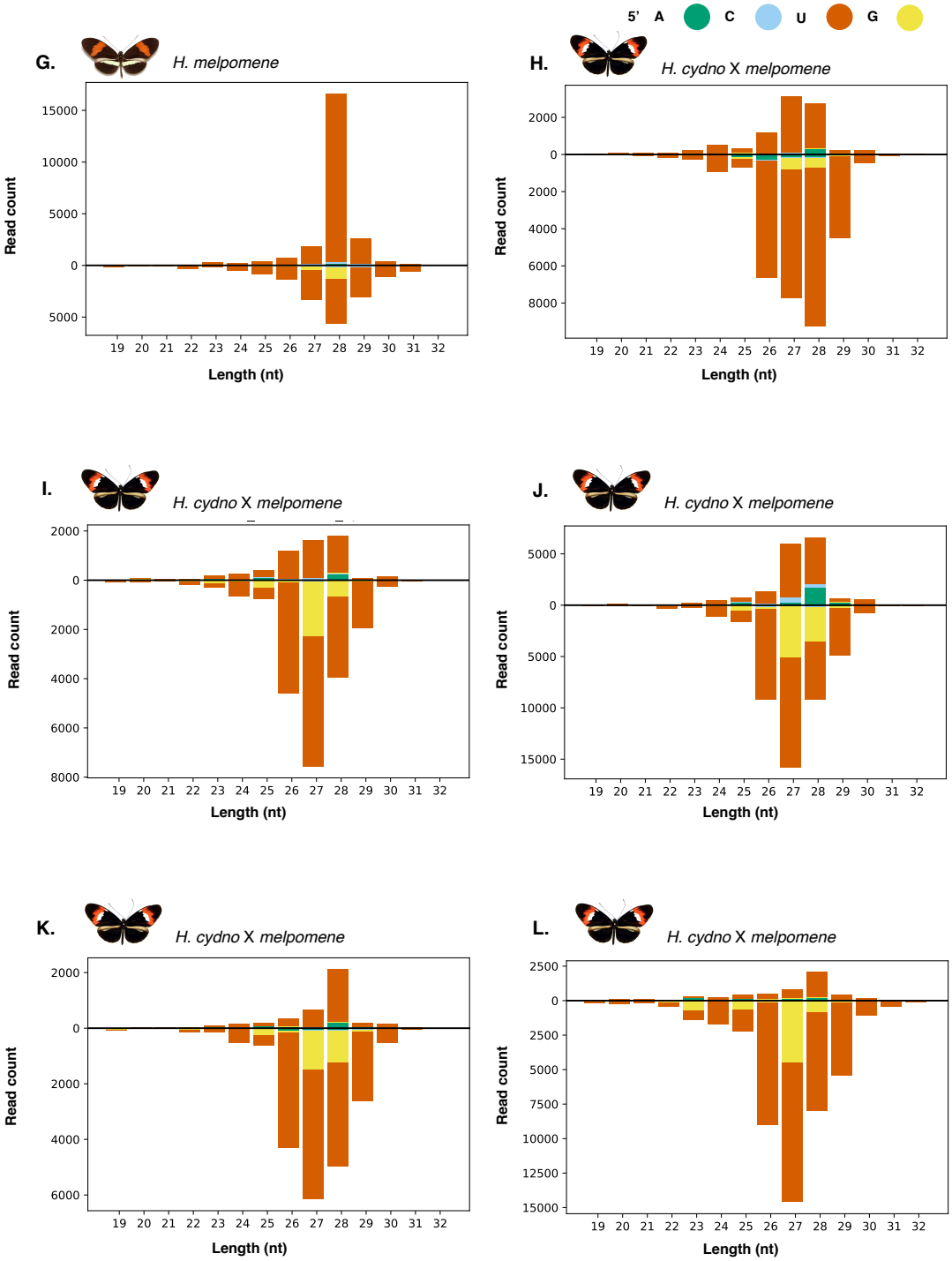
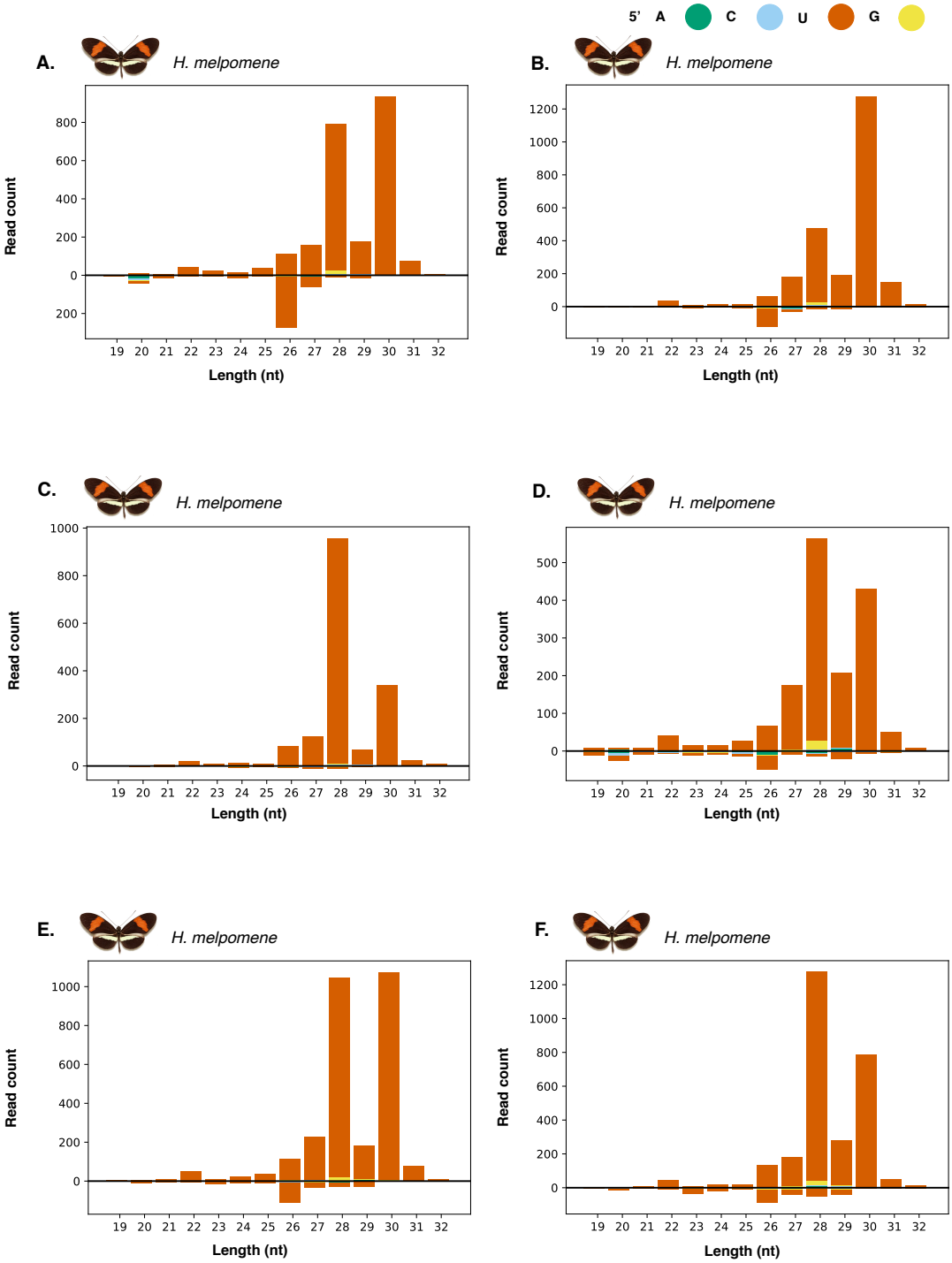

5' A C U G
G.
H. melpomene
H.
H. cydno X melpomene
I.
H. cydno X melpomene
J.
H. cydno X melpomene
K.
H. cydno X melpomene
L.
H. cydno X melpomene
Read count
Length (nt)

Figure S32. *H. melpomene* and hybrid sRNAs mapping to *H. melpomene*'s nPIF-5 TE

sRNA read distribution for *H. melpomene* and *H. cydno* x *H. melpomene* samples mapped to *H. melpomene*'s nPIF-5 TE. *H. melpomene*'s nPIF-5 TE is significantly more abundant in hybrid samples than *H. melpomene* samples (Figure 4, Table 2). y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. *H. melpomene* samples: **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33. *H. cydno* x *H. melpomene* samples: **H.** Sample AP50, **I.** Sample AP57, **J.** Sample AP59, **K.** Sample AP60, and **L.** Sample AP72.


5' A C U G
A.
H. melpomene
B.
H. melpomene
C.
H. melpomene
D.
H. melpomene
E.
H. melpomene
F.
H. melpomene
Read count
Length (nt)

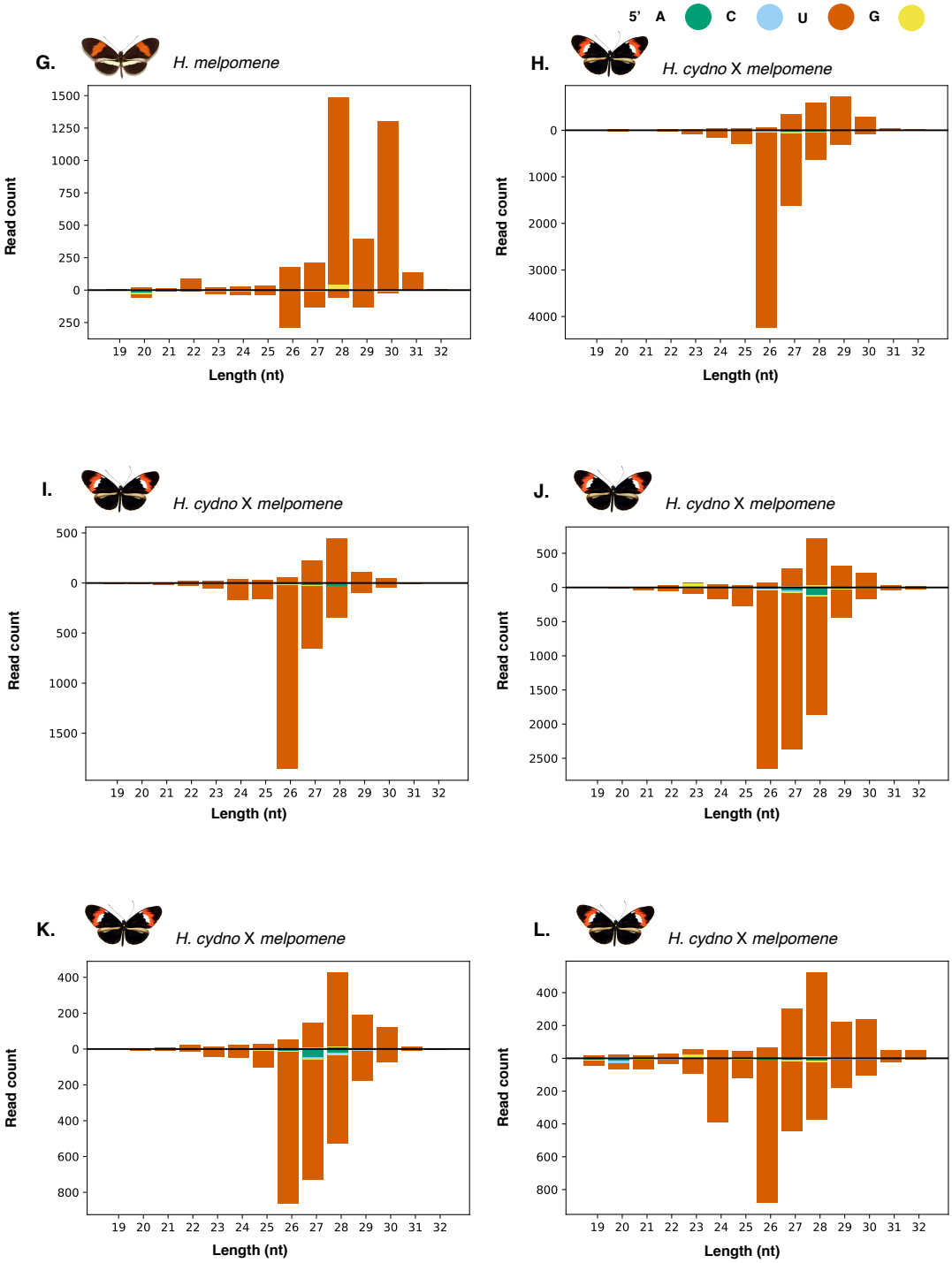

5' A C U G
G.
H. melpomene
H.
H. cydno X melpomene
Read count
Length (nt)
I.
H. cydno X melpomene
J.
H. cydno X melpomene
K.
H. cydno X melpomene
L.
H. cydno X melpomene

Figure S33. *H. melpomene* and hybrid sRNAs mapping to *H. melpomene*'s Polinton-2 TE

sRNA read distribution for *H. melpomene* and *H. cydno* x *H. melpomene* samples mapped to *H. melpomene*'s Polinton-2 TE. *H. melpomene*'s Polinton-2 TE is significantly more abundant in hybrid samples than *H. melpomene* samples (Figure 4, Table 2). y axis: read count, x axis: length of the clean reads (nucleotides). Colour: 5' nucleotide of the read. *H. melpomene* samples: **A.** Sample AP2, **B.** Sample AP4, **C.** Sample AP13, **D.** Sample AP16, **E.** Sample AP22, **F.** Sample AP30, and **G.** Sample AP33. *H. cydno* x *H. melpomene* samples: **H.** Sample AP50, **I.** Sample AP57, **J.** Sample AP59, **K.** Sample AP60, and **L.** Sample AP72.

"We feel to be as near witnesses, as we can ever hope to be, of the creation of a new species on this earth"

After reading the descriptions by Henry Walter Bates of *Heliconius* species (Bates, 1862), Charles Darwin wrote: "It is hardly an exaggeration to say, that whilst reading and reflecting on the various facts given in this Memoir, we feel to be as near witnesses, as we can ever hope to be, of the creation of a new species on this earth" (Darwin, 1863). Charles Darwin was fascinated not only by the diversity of butterfly species in the tropics, but also by their wing pattern diversity and the mimicry rings they formed. Since then, the evolution and genetics of wing colour patterns have been the main focus of *Heliconius* research with many *Heliconius* races diverging only at these genomic regions (Nadeau *et al.*, 2011; Martin *et al.*, 2013). However, with a known genetic basis for species incompatibilities, species pairs at different levels of divergence, and inter-specific hybrids with varying degrees of fitness, *Heliconius* is an excellent system to investigate the genetic basis of barriers to interspecific gene flow.

E. O. Wilson defines species as the fundamental unit of biodiversity. He argues that, not having a natural unit such as species, would be abandoning "obvious entities" and concede to an idea of "amorphous variation" with "arbitrary limits" (Wilson, 1992). Like E. O. Wilson, some biologists believe that species are objectively identifiable and that genera or subspecies do not have the same logical precision (Mallet, 1998). Charles Darwin, however, did not share this opinion and believed that species were not more *logical* than other taxonomic levels (Darwin, 1859). Species are instead fundamental units

of *local* biodiversity and the clarity associated with these “obvious entities” becomes less meaningful with time and space (Mallet, 1998). Species are usually ecologically distinct to be able to coexist. Behaviour, morphological or genetic variation that differentiates species is likely to be both ecologically and evolutionary significant. However, distinction between highly differentiated forms of a group of individuals that are geographically isolated becomes less clear. For example, differentiating between mallard (*Anas platyrhynchos*) populations distributed worldwide is not straightforward, and how many mallard populations are *true* species is open to debate (Mallet 1998).

Agreement over distinctions between species and lower taxonomic levels has often been elusive (Merrill *et al.*, 2015). The *Biological Species Concept* defines species as “a population whose members are able to interbreed freely under natural conditions” and so, reproductive isolation, is explicitly considered to underlie the species barrier (Mayr 1942). However, the strength of reproductive isolation observed between species is broadly continuous, and so, the degree of reproductive isolation required for species status, is not always obvious (Presgraves, 2002; Mallet, 2007; Merrill *et al.*, 2011; Crespi and Nosil, 2013). Another challenge to the *Biological Species Concept* is the fact that hybridization among species is relatively common on a per-species basis. Approximately 10-30% of multicellular animals and plants hybridize (Abbott *et al.*, 2013). Gene exchange is “widespread and substantial between sympatric taxa” (Coyne and Orr, 2004). Among those that hybridize, between 1 in 100 and 1 in 10 000 individuals are hybrids when in sympatry (Mallet, 2005).

To delineate species barriers I favour the *Genotypic Cluster Definition* of species over the *Biological Species Concept* (Mallet, 1995). The *Genotypic Cluster Definition* uses sympatric coexistence of distinct multilocus genotypes as the defining character of species. Divergence and maintenance of different species in sympatry has a genomic architecture that is distinct at multiple loci, and the loci involved in reproductive isolation are therefore necessarily involved in the ongoing maintenance of species during speciation (Mallet,

1995). For example, *H. erato* and *H. himera* are classified as distinct species because where they co-occur hybrids are rare (Jiggins *et al*., 1996). Species, as defined by the *Genotypic Cluster Definition*, are characterised by a bimodal distribution of traits even if gene flow exists.

Whole genome re-sequencing data of *Heliconius* butterfly samples fits the *Genotypic Cluster Definition* of species. *H. cydno* and *H. melpomene*, for example, have high levels of genome-wide divergence despite occasional hybridisation and genome-wide signatures of admixture (Martin *et al*., 2013; Kronforst *et al*., 2013). The gene expression data from *H. cydno* and *H. melpomene* I analysed further illustrates this, and there are significant gene expression differences between *H. cydno* and *H. melpomene* ovary tissue (Chapter 3, "Sterility in *Heliconius cydno* x *Heliconius melpomene* F1 female hybrids: a phenotypic and gene expression study of hybrid incompatibilities"). Moreover, I have shown for the first time in *Heliconius*, that there are also differences at the non-coding transcript level between the two species (Chapter 4, "piRNA mediated epigenetic silencing does not underlie post-zygotic isolation between *Heliconius cydno* and *Heliconius melpomene*").

There is genome-wide divergence at the level of nucleotide composition (Martin *et al*., 2013; Kronforst *et al*., 2013) and coding and non-coding expression between *H. cydno* and *H. melpomene*. However, divergence between different *H. melpomene* races, is only present at a few loci under divergent selection (The Heliconius Genome Consortium, 2012; Martin *et al*., 2013; 2016). Through the emphasis of multilocus genotypes, the *Genotypic Cluster Definition* is a useful tool for investigating the maintenance of distinct species, and the genomic architecture of gene flow and divergence in sympatry.

Hybridization is reproduction between members of genetically distinct populations that produces offspring of mixed ancestry, and it occurs in almost all processes of speciation (Barton and Hewitt, 1989). Patterns of contemporary hybridization are a snapshot complex interactions and the

evolution of complete reproductive isolation might take hundreds to millions of generations (Abbott *et al.*, 2013). Therefore, studies concerning the outcomes and significance of hybridization, need to consider the relative spatial and temporal context where it is occurring.

H. cydno and *H. melpomene* are sister species living in sympatry that differ in many aspects of their ecology and behaviour (Jiggins *et al.*, 2008). Hybridization between *H. cydno* and *H. melpomene* always results in sterile F1 females (Naisbit *et al.*, 2002) and there is strong disruptive selection against the hybrid colour patterns (Merrill *et al.*, 2012). Rapid ecological divergence seems to be a driver of the earliest stages of speciation in *Heliconius* and so, barriers to gene flow between *H. cydno* and *H. melpomene* might have accumulated during periods of spatial isolation or due to other obstacles to dispersal (McMillan *et al.*, 1997; Jiggins *et al.*, 2001; Muñoz *et al.*, 2010; Abbott *et al.*, 2013). Regardless of how the barriers accumulated, barrier loci between *H. cydno* and *H. melpomene* are either under divergent selection or contribute to reduced hybrid fitness or assortative mating.

Coupling between loci that contribute to isolation between *H. cydno* and *H. melpomene* is expected to build up through evolutionary time depending on the overall antagonism between selection and recombination among diverging loci (Kruuk *et al.*, 1999). Explicitly addressing the mechanisms that result in reproductive isolation between *H. cydno* and *H. melpomene* is one way to study species identity and diversity. Studying the genetics of inter-species *Heliconius* crosses and identifying regions of exceptional divergence between two species can elucidate the architecture of reproductive isolation and, ultimately, the process of speciation.

Different colour patterns between *H. cydno* and *H. melpomene*, and the associated shift in mimicry rings, contribute to speciation of these two species but need to be associated with divergent mate preference to cause assortative mating. Genetic coupling of colour pattern and mate preference loci contributes to speciation as it results in progressively independent

evolutionary trajectories of both species; eventually resulting in the build-up of reproductive barriers through linkage disequilibrium between adaptive and assortative mating loci (Felsenstein, 1981; Jiggins *et al.*, 2001; Merrill *et al.,* 2011). The genetic basis for wing pattern and mate preference between *H. cydno* and *H. melpomene* is controlled by a few loci of major effect (Merrill *et al.*, 2011; Reed *et al.*, 2011; Wallbank *et al.*, 2016). However, despite these loci of large effect that differ between *H. cydno* and *H. melpomene*, across the genome, there is evidence for pervasive polygenic selection maintaining species differences with some of these loci likely to be reproductive barriers (Martin *et al.*, 2016; Martin pers. comm.; Roux pers. comm.).

Structural genomic differences between two species are expected to contribute to barriers to gene flow (Noor *et al.*, 2001). Recombination suppression is unlikely to contribute to speciation between *H. cydno* and *H. melpomene* as there is no evidence of chromosomal inversions between the two sympatric species (Davey *et al.* 2017). I focused on duplications, another type of structural variation, between *H. cydno* and *H. melpomene* and identified duplicated loci putatively under selection that have a potential role in host plant and mate recognition differences (Chapter 1, “The comparative landscape of duplication in *Heliconius melpomene* and *Heliconius cydno*”). The duplicated loci I identified, as well as the distribution of differentially expressed genes between *H. cydno*, *H. melpomene* and the F1 *H. cydno* x *H. melpomene* hybrids (Chapter 3, “Sterility in *Heliconius cydno* x *Heliconius melpomene* F1 female hybrids: a phenotypic and gene expression study of hybrid incompatibilities”) are distributed throughout the genome. The pervasive distribution of both duplicated loci and differentially expressed genes between *H. cydno* and *H. melpomene* serves as further evidence for a role of polygenic selection maintaining species boundaries and does not support a scenario of a few islands of differentiation flanking positively selected loci.

Transposable element (TE) distribution can impact recombination rate and reproductive compatibility between two hybridizing species. After merging two

divergent genomes in F1 hybrids there may be quantitative and qualitative mismatches between TEs and the maternally inherited sRNAs (Kidwell, 1983; Kelleher *et al.*, 2012; Czech and Hannon, 2016). Mis-regulation of TEs can lead to epigenetic re-patterning throughout the hybrid genome and activation of certain TEs. Foreign transposon families can dramatically reduce the fitness of the new host individual or host population as observed in *Drosophila* (i.e. hybrid dygenesis) (Brennecke *et al.*, 2008).

Limited transposition, however, might also result in structural polymorphisms and recombination rate changes (Dooner and He, 2008; Witherspoon *et al.*, 2009). TE element mis-expression can be a driver of the speciation process because it might: 1) trigger genome-wide variation in functional genes (Wang *et al.*, 2013); or 2) modify recombination patterns across the genome (Michalak, 2009). From previous studies we know that recombination rates in the *H. cydno* x *H. melpomene* hybrids does not appear to differ from the parents and so, a mechanism like hybrid dygenesis was unlikely to underlie reproductive isolation between the two species (Davey *et al.,* 2017). However, we did not know whether the sRNA pathway was functional in the hybrids or whether TEs were mis-expressed. In Chapter 4, "piRNA mediated epigenetic silencing does not underlie post-zygotic isolation between *Heliconius cydno* and *Heliconius melpomene*", I have shown that the piRNA pathway and TE de-repression are not correlated to hybrid sterility. This is interesting once over three decades of *Drosophila* hybrid sterility research have successively identified epigenetic mechanisms as the main cause of reproductive isolation between closely related *Drosophila* species or strains. Here, I show that the piRNA pathway is functional in the hybrids contrary to my initial assumption, which leads to the question of how fast may the piRNA pathway genes evolve. Studies in *Drosophila* have shown that they evolve faster than microRNAs but slower than small interfering RNAs and it will be interesting to test if this is also the case for *Heliconius* (Obbard, Gordon, *et al.*, 2009; Obbard, Welch, *et al.*, 2009).

Functional changes can affect every aspect of the hybrid phenotype and can have a role in both the establishment of barriers that reduce fitness or, alternatively, in generating evolutionary novelty. Hybrid traits that reduce or increase fitness tend to be defined as qualitatively different, however, both relate to the appearance of fitness-related phenotypic traits in hybrids that lie outside the parent's distribution. Sex-biased genes have been shown to evolve faster than unbiased genes and may be disproportionally involved in the maintenance of species barriers. For example reproductive tract proteins are amongst the fastest evolving in the *Drosophila* genome (Parisi *et al.*, 2004; Panhuis *et al.*, 2006; Singh and Jagadeeshan, 2012). In Chapter 2, "Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*", I identified sex-biased genes and show that these are unlikely to play a disproportionate role in maintaining species barriers as they do not have significantly faster rates of evolution compared to the transcriptome average.

For most areas of evolutionary biology, population genetics theory has precluded the technical developments to experimentally test the accuracy of its predictions. Faster-sex chromosome evolution is one of such areas of molecular evolution. With whole genome sequencing technology we can finally test empirically for faster-sex chromosome divergence or adaptation in a variety of taxa. Sex-chromosome to autosome substitution rates contributing to total divergence between the two, result not only from adaptive but also from neutral and slightly deleterious substitutions. In lineages with large effective population sizes, positive selection is more efficient, predicting faster-sex chromosome evolution, but purifying evolution is also more efficient. The mating system, standing genetic variation, and dosage compensation mechanisms can also affect the overall rate of adaptive substitutions in the sex chromosome.

In *Heliconius* I did not find a fast-Z effect (Chapter 2, "Lack of the fast-Z effect: sexually dimorphic expression and transcriptome evolution in *Heliconius melpomene*"). In contrast to what has been reported for *Drosophila* where

large N_{eX}/N_{eA} is likely to be one of the drivers of faster-X adaptation, in *Heliconius,* males have greater variance in reproductive success and so there is a low N_{eX}/N_{eZ}. A low N_{eX}/N_{eZ}, coupled with the fact that there is not a complete mechanism of dosage compensation, may explain the lack of a fast-Z effect in *Heliconius* where evolution from standing genetic variation might also reduce the opportunity for faster-Z evolution. In the future, it would be interesting to explore this by establishing whether there is a relationship between gene expression level and chromosome type in *Heliconius* (Rousselle *et al.*, 2016). In primates, for example, non-synonymous substitutions are negatively correlated with expression level and so this can be interpreted as an increased strength of purifying selection on highly expressed genes (Nguyen *et al.*, 2015). Performing such analysis will allow a clearer understanding of chromosome evolution in female heterogametic taxa, which ultimately clarifies the genomic architecture of divergence in such species.

A fundamental question in evolutionary biology regards the genetic mechanisms underlying speciation. Different evolutionary forces, acting within a certain population genetics environment, drive the genomic basis of species differences. Through the study of these evolutionary forces and the environment they occupy, we further our understanding of the genetics of speciation. By studying the genetics of *H. cydno* and *H. melpomene* inter-specific hybrids I have tested some of the possible mechanisms underlying the origin of genomic incompatibilities between two species. I show that there are significant gene expression differences between *H. cydno* and *H. melpomene* ovary tissues and map features that may confer advantageous adaptations or that could be important in preserving each species. It is important to note that the crosses I used in this study were not inbred and hence, they will confer an accurate representation of the natural and genetic diversity in *H. cydno* and *H. melpomene.* However, simultaneously, the crosses used here will also not capture all the natural diversity in *H. cydno* and *H. melpomene*. Some of the conclusions might simply portrait the

mechanisms and process I was able capture from the specific samples analysed. Finally, genomic tools have allowed us to start to truly appreciate the evolutionary importance of hybridization, and further studies in *Heliconius* and other systems will continue to increase our understanding of speciation in the face of gene flow.

REFERENCES

Abbott JK, Nordén AK, Hansson B (2017). Sex chromosome evolution: historical insights and future perspectives. Proc Biol Sci 284: 20162806.

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, et al. (2013). Hybridization and speciation. Journal of Evolutionary Biology 26: 229–246.

Abyzov A, Urban AE, Snyder M, Gerstein M (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research 21: 974–984.

Anders S, Pyl PT, Huber W (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166–169.

Anderson LK, Covey PA, Larsen LR, Bedinger P, Stack SM (2010). Structural differences in chromosomes distinguish species in the tomato clade. Cytogenet Genome Res 129: 24–34.

Armengol L, Villatoro S, González JR, Pantano L, García-Aragonés M, Rabionet R, et al. (2009). Identification of copy number variants defining genomic differences among major human groups. (M Bauchet, Ed.). PLoS ONE 4: e7230.

Arroyo JI, Hoffmann FG, Opazo JC (2012). Gene duplication and positive selection explains unusual physiological roles of the relaxin gene in the European rabbit. J Mol Evol 74: 52–60.

Ashe A, Bélicard T, Le Pen J, Sarkies P, Frezal L, Lehrbach NJ, et al. (2013). A deletion polymorphism in the Caenorhabditis elegans RIG-I homolog disables viral RNA dicing and antiviral immunity. Elife 2: e00994.

Assis R, Zhou Q, Bachtrog D (2012). Sex-biased transcriptome evolution in Drosophila. Genome Biology and Evolution 4: 1189–1200.

Avila V, Campos JL, Charlesworth B (2015). The effects of sex-biased gene expression and X-linkage on rates of adaptive protein sequence evolution in Drosophila. Biology Letters 11: 20150117–20150117.

Baines JF, Harr B (2007). Reduced X-linked diversity in derived populations of house mice. Genetics 175: 1911–1921.

Bartolomé C, Bello X, Maside X (2009). Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes. Genome Biol 10: R22.

Barton NH (1995). A general model for the evolution of recombination. Genet Res 65: 123–145.

Barton NH, Hewitt GM (1989). Adaptation, speciation and hybrid zones. Nature 341: 497–503.

Bates HW (1862). Contributions to an insect fauna of the Amazon valley. Lepidoptera: Heliconidae. Transactions of the Linnean Society of London. 23: 495-566.

Bateson W (1909). Heredity and variation in modern lights. In: Darwin and Modern Science (A. C. Seward, ed.), pp. 85–101. Cambridge University Press, Cambridge.

Begun DJ, Aquadro CF (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356: 519–520.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, et al. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. (MAF Noor, Ed.). PLoS Biol 5: e310.

Beisswanger S, Stephan W (2008). Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in Drosophila. Proc Natl Acad Sci USA 105: 5447–5452.

Bell LR, Maine EM, Schedl P, Cline TW (1988). Sex-lethal, a Drosophila sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. Cell 55: 1037–1046.

Biddle FG, Eales BA, Dean WL (1994). Haldane's rule and heterogametic female and male sterility in the mouse. Genome 37: 198–202.

Bikard D, Patel D, Le Metté C, Giorgi V, Camilleri C, Bennett MJ, et al. (2009). Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. Science 323: 623–626.

Blackman RK, Grimaila R, Koehler MM, Gelbart WM (1987). Mobilization of hobo elements residing within the decapentaplegic gene complex: suggestion of a new hybrid dysgenesis system in Drosophila melanogaster. Cell 49: 497–505.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE (2012). Genomic analysis of a key innovation in an experimental Escherichia coli population. Nature 489: 513–518.

Bolnick DI, Nosil P (2007). Natural selection in populations subject to a migration load. Evolution 61: 2229–2243.

Bourc'his D, Bestor TH (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431: 96–99.

Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ (2008). An epigenetic role for maternally inherited piRNAs in transposon silencing. Science 322: 1387–1392.

Brideau NJ, Flores HA, Wang J, Maheshwari S, Wang X, Barbash DA (2006). Two Dobzhansky-Muller genes interact to cause hybrid lethality in

Drosophila. Science 314: 1292–1295.

Briscoe AD, Macias-Muñoz A, Kozak KM, Walters JR, Yuan F, Jamie GA, et al. (2013). Female behaviour drives expression and evolution of gustatory receptors in butterflies. (J Zhang, Ed.). PLoS Genet 9: e1003620.

Brookfield JF (1986). The population biology of transposable elements. Philosophical Transactions of the Royal Society B: Biological Sciences 312: 217–226.

Brothers AN, Delph LF (2010). Haldane's rule is extended to plants with sex chromosomes. Evolution 64: 3643–3648.

Bucheton A, Lavige JM, Picard G, L'Heritier P (1976). Non-mendelian female sterility in Drosophila melanogaster: quantitative variations in the efficiency of inducer and reactive strains. Heredity 36: 305–314.

Camus MF, Wolf JBW, Morrow EH, Dowling DK (2015). Single Nucleotides in the mtDNA Sequence Modify Mitochondrial Molecular Function and Are Associated with Sex-Specific Effects on Fertility and Aging. Curr Biol 25: 2717–2722.

Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar JA, et al. (2012). Evidence for widespread positive and purifying selection across the European rabbit (Oryctolagus cuniculus) genome. Molecular Biology and Evolution 29: 1837–1849.

Carracedo MC, Suarez A, Asenjo A, Casares P (1998). Genetics of hybridization between Drosophila simulans females and D. melanogaster males. Heredity 80 (Pt 1): 17–24.

Carter AJR, Hermisson J, Hansen TF (2005). The role of epistatic gene interactions in the response to selection and the evolution of evolvability. Theor Popul Biol 68: 179–196.

Carvalho AB, Clark AG (2013). Efficient identification of Y chromosome

sequences in the human and Drosophila genomes. Genome Research 23: 1894–1907.

Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, Kalbe M, et al. (2014). Extensive copy-number variation of young genes across stickleback populations. (J Zhang, Ed.). PLoS Genet 10: e1004830.

Challis RJ, Kumar S, Dasmahapatra KKK, Jiggins CD, Blaxter M. Lepbase: the Lepidopteran genome database. bioRxiv 056994; doi: https://doi.org/10.1101/056994

Charlesworth (2002). The Genetics and Biology of Sex Determination. John Wiley & Sons. Editors Chadwick DJ and Goode JA.

Charlesworth B (1987). The population biology of transposable elements. Trends in Ecology & Evolution 2: 21–23.

Charlesworth B (1996a). The evolution of chromosomal sex determination and dosage compensation. Curr Biol 6: 149–162.

Charlesworth B (1996b). Background selection and patterns of genetic diversity in Drosophila melanogaster. Genet Res 68: 131–149.

Charlesworth B (2001). The effect of life-history and mode of inheritance on neutral genetic variability. Genet Res 77: 153–166.

Charlesworth B (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10: 195–205.

Charlesworth B (2012). The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the Drosophila X chromosome. Genetics 191: 233–246.

Charlesworth B, Coyne JA, and Barton NH (1987). The relative rates of evolution of sex chromosomes and autosomes. The American Naturalist 130:1, 113-146.

Cheang CC, Tsang LM, Chu KH, Cheng I-J, Chan BKK (2013). Host-specific phenotypic plasticity of the turtle barnacle Chelonibia testudinaria: a widespread generalist rather than a specialist. (V Laudet, Ed.). PLoS ONE 8: e57592.

Chen K, Chen L, Fan X, Wallis J, Ding L, Weinstock G (2014). TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. Genome Research 24: 310–317.

Chuong EB, Elde NC, Feschotte C (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet 18: 71–86.

Clark RM, Wagler TN, Quijada P, Doebley J (2006). A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet 38: 594–597.

Combes M-C, Hueber Y, Dereeper A, Rialle S, Herrera J-C, Lashermes P (2015). Regulatory divergence between parental alleles determines gene expression patterns in hybrids. Genome Biology and Evolution 7: 1110–1121.

Conant GC, Wolfe KH (2008). Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9: 938–950.

Connallon T, Singh ND, Clark AG (2012). Impact of genetic architecture on the relative rates of X versus autosomal adaptive substitution. Molecular Biology and Evolution 29: 1933–1942.

Conrad DF, Hurles ME (2007). The population genetics of structural variation. Nat Genet 39: S30–6.

Coyne JA (1985). The genetic basis of Haldane's rule. Nature 314: 736–738.

Coyne JA and Orr HA (2004) Speciation. Sinauer Associates, Inc., Sunderland, Massachusetts.

Coyne JA, Orr HA (1997). Patterns of Speciation in *Drosophila* revisited. Evolution 51: 295–303.

Coyne JA, Orr HA (1998). The evolutionary genetics of speciation. Philosophical Transactions of the Royal Society B: Biological Sciences 353: 287–305.

Coyne, J.A. & Orr, H.A. 2004. Speciation, 1st edn. Sinauer Associates Inc, Sunderland, MA.

Crawley MJ (2005) Statistics: An Introduction using R. Wiley, Chichester.

Crespi B, Nosil P (2013). Conflictual speciation: species formation via genomic conflict. Trends in Ecology & Evolution 28: 48–57.

Cruickshank TE, Hahn MW (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular Ecology 23: 3133–3157.

Csankovszki G, McDonel P, Meyer BJ (2004). Recruitment and spreading of the C. elegans dosage compensation complex along X chromosomes. Science 303: 1182–1185.

Czech B, Hannon GJ (2016). One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem Sci 41: 324–337.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks et al., 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

Darwin C. (1863) [Review of] Contributions to an insect fauna of the Amazon Valley. By Henry Wlater Bates, Esq. Transactions of the Linnean Society. Vol XXIII. 1862, p495. Natual History Review. 3: 219-224

Darwin C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, 1st ed. John Murray, London.

Davey JW, Barker SL, Rastas PM, Pinharanda A, Martin SH, Durbin R, et al. (2017). No evidence for maintenance of a sympatric Heliconius species barrier by chromosomal inversions. Evolution Letters: n/a–n/a.

Davey JW, Chouteau M, Barker SL, Maroja L, Baxter SW, Simpson F, et al. (2016). Major Improvements to the Heliconius melpomene Genome Assembly Used to Confirm 10 Chromosome Fusion Events in 6 Million Years of Butterfly Evolution. G3: GenesIGenomesIGenetics: g3.115.023655.

Davies N (1996). Haldane‘s rule is dead, long live Haldane’s rule. Trends in Ecology & Evolution 11: 508.

De Mita S, Siol M (2012). EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet 13: 27.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Publishing Group 43: 491–498.

Deng C, Cheng C-HC, Ye H, He X, Chen L (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proc Natl Acad Sci USA 107: 21593–21598.

Ding Y, Zhao L, Yang S, Jiang Y, Chen Y, Zhao R, et al. (2010). A young Drosophila duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. (J Zhang, Ed.). PLoS Genet 6: e1001255.

Dion-Côté A-M, Renaut S, Normandeau E, Bernatchez L (2014). RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. Molecular Biology and Evolution 31: 1188–1199.

Dobzhansky T (1937). Genetics and the Origin of Species. Columbia University Press: New York.

Dobzhansky T (1936). Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. Genetics 21: 113–135.

Dooner HK, He L (2008). Maize genome structure variation: interplay between retrotransposon polymorphisms and genic recombination. Plant Cell 20: 249–258.

Dorus S, Gilbert SL, Forster ML, Barndt RJ, Lahn BT (2003). The CDY-related gene family: coordinated evolution in copy number, expression profile and protein sequence. Hum Mol Genet 12: 1643–1650.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203–218.

Dulai KS, Dornum von M, Mollon JD, Hunt DM (1999). The evolution of trichromatic color vision by opsin gene duplication in New World and Old World primates. Genome Research 9: 629–638.

Dunlap-Pianka H, Boggs CL, Gilbert LE (1977). Ovarian Dynamics in Heliconiine Butterflies: Programmed Senescence versus Eternal Youth. Science 197: 487–490.

Dutheil J, Boussau B (2008). Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evolutionary Biology 8: 255.

Duvaux L, Geissmann Q, Gharbi K, Zhou J-J, Ferrari J, SMADJA CM, et al. (2015). Dynamics of copy number variation in host races of the pea aphid. Molecular Biology and Evolution 32: 63–80.

Ellegren H, Galtier N (2016). Determinants of genetic diversity. Nat Rev Genet 17: 422–433.

Ellegren H, Parsch J (2007). The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet 8: 689–698.

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, et al. (2012). The genomic landscape of species divergence in Ficedula flycatchers. Nature 491: 756–760.

Emelianov I, Marec F, Mallet J (2004). Genomic evidence for divergence with gene flow in host races of the larch budmoth. Proceedings of the Royal Society B: Biological Sciences 271: 97–105.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science 320: 1629–1631.

Farslow JC, Lipinski KJ, Packard LB, Edgley ML, Taylor J, Flibotte S, et al. (2015). Rapid Increase in frequency of gene copy-number variants during experimental evolution in Caenorhabditis elegans. BMC Genomics 16: 1044.

Felsenstein J (1981). Skepticism towards Santa Rosalia, or why there are there so few kinds of animals? Evolution 35: 124–138.

Feuk L, Carson AR, Scherer SW (2006). Structural variation in the human genome. Nat Rev Genet 7: 85–97.

Feulner PGD, Chain FJJ, Panchal M, Eizeguirre C, Kalbe M, Lenz TL, et al. (2013). Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. Molecular Ecology 22: 635–649.

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. (2016). Detection of human adaptation during the past 2000 years. Science 354: 760–764.

Foll M, Gaggiotti O (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180: 977–993.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999).

Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.

Galli M, Theriault A, Liu D, Crawford NM (2003). Expression of the Arabidopsis transposable element Tag1 is targeted to developing gametophytes. Genetics 165: 2093–2105.

Galtier N, Bazin E, Bierne N (2006). GC-biased segregation of noncoding polymorphisms in Drosophila. Genetics 172: 221–228.

Gante HF, Matschiner M, Malmstrøm M, Jakobsen KS, Jentoft S, Salzburger W (2016). Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. Molecular Ecology 25: 6143–6161.

García-Ramos G, Kirkpatrick M (1997). Genetic models of adaptation and gene flow in peripheral populations. Evolution 51: 21–28.

Gautier M (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics 201: 1555–1579.

Gilbert N, Lutz-Prigge S, Moran JV (2002). Genomic deletions created upon LINE-1 retrotransposition. Cell 110: 315–325.

Gourbière S, Mallet J (2010). Are species real? The shape of the species boundary with exponential failure, reinforcement, and the "missing snowball". Evolution 64: 1–24.

Grath S, Parsch J (2016). Sex-Biased Gene Expression. Annu Rev Genet 50: 29–44.

Grentzinger T, Armenise C, Brun C, Mugat B, Serrano V, Pelisson A, et al. (2012). piRNA-mediated transgenerational inheritance of an acquired trait. Genome Research 22: 1877–1888.

Haerty W, Singh RS (2006). Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of

Drosophila. Molecular Biology and Evolution 23: 1707–1714.

Hahn MW, Han MV, Han S-G (2007). Gene family evolution across 12 Drosophila genomes. PLoS Genet 3: e197.

Haldane, J. B. S. (1922). "Sex ratio and unisexual sterility in hybrid animals". J. Genet. 12: 101–109. doi:10.1007/BF02983075.

Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD (2008). Sex-biased evolutionary forces shape genomic patterns of human diversity. (DA Petrov, Ed.). PLoS Genet 4: e1000202.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009). Mechanisms of change in gene copy number. Nat Rev Genet 10: 551–564.

Hegarty MJ, Barker GL, Brennan AC, Edwards KJ, Abbott RJ, Hiscock SJ (2008). Changes to gene expression associated with hybrid speciation in plants: further insights from transcriptomic studies in Senecio. Philos Trans R Soc Lond, B, Biol Sci 363: 3055–3069.

Hill T, Schlötterer C, Betancourt AJ (2016). Hybrid Dysgenesis in Drosophila simulans Associated with a Rapid Invasion of the P-Element. (HS Malik, Ed.). PLoS Genet 12: e1005920.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, et al. (2002). The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129–149.

Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee Apis mellifera. Nature 443: 931–949.

Hu TT, Eisen MB, Thornton KR, Andolfatto P (2013). A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. Genome Research 23: 89–98.

Hunt DM, Dulai KS, Cowing JA, Julliot C, Mollon JD, Bowmaker JK, et al.

(1998). Molecular evolution of trichromacy in primates. Vision Res 38: 3299–3306.

Innocenti P, Morrow EH, Dowling DK (2011). Experimental evidence supports a sex-specific selective sieve in mitochondrial genome evolution. Science 332: 845–848.

International Chicken Genome Sequencing Consortium (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695–716.

Irish VF, Litt A (2005). Flower development and evolution: gene duplication, diversification and redeployment. Current Opinion in Genetics & Development 15: 454–460.

Iskow RC, Gokcumen O, Lee C (2012). Exploring the role of copy number variants in human adaptation. Trends in Genetics 28: 245–257.

Jakšić AM, Kofler R, Schlötterer C (2017). Regulation of transposable elements: Interplay between TE-encoded regulatory sequences and host-specific trans-acting factors in Drosophila melanogaster. Molecular Ecology 26: 5149–5159.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004). Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573.

Jiggins CD, McMillan WO, Neukirchen W, Mallet J, mall (1996). What can hybrid zones tell us about speciation? The case of Heliconius erato and H. himera (Lepidoptera: Nymphalidae). : 1–22.

Jiggins CD, Naisbit RE, Coe RL, Mallet J (2001). Reproductive isolation caused by colour pattern mimicry. Nature 411: 302–305.

Jiggins CD, Salazar C, Linares M, Mavarez J (2008). Hybrid trait speciation and Heliconius butterflies. Philosophical Transactions of the Royal Society

B: Biological Sciences 363: 3047–3054.

Jiggins CD, Salazar C, Linares M, Mavarez J (2008). Hybrid trait speciation and Heliconius butterflies. Philosophical Transactions of the Royal Society B: Biological Sciences 363: 3047–3054.

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, et al. (2001). Positive selection of a gene family during the emergence of humans and African apes. Nature 413: 514–519.

Jombart T, Ahmed I (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27: 3070–3071.

Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, et al. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature 477: 203–206.

Josephs EB, Wright SI (2016). On the Trail of Linked Selection. (NH Barton, Ed.). PLoS Genet 12: e1006240.

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA (2009). Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. Genome Research 19: 1404–1418.

Katju V (2012). In with the old, in with the new: the promiscuity of the duplication process engenders diverse pathways for novel gene creation. International Journal of Evolutionary Biology 2012: 341932–24.

Katju V, Bergthorsson U (2013). Copy-number changes in evolution: rates, fitness effects and adaptive significance. Front Genet 4: 273.

Kawakami T, Dhakal P, Katterhenry AN, Heatherington CA, Ungerer MC (2011). Transposable element proliferation and genome expansion are rare in contemporary sunflower hybrid populations despite widespread transcriptional activity of LTR retrotransposons. Genome Biology and Evolution 3: 156–167.

Keane OM, Toft C, Carretero-Paulet L, Jones GW, Fares MA (2014). Preservation of genetic and regulatory robustness in ancient gene duplicates of Saccharomyces cerevisiae. Genome Research 24: 1830–1841.

Kelleher ES, Edelman NB, Barbash DA (2012). Drosophila interspecific hybrids phenocopy piRNA-pathway mutants. (MAF Noor, Ed.). PLoS Biol 10: e1001428.

Khanduja JS, Calvo IA, Joh RI, Hill IT, Motamedi M (2016). Nuclear Noncoding RNAs and Genome Stability. Mol Cell 63: 7–20.

Kidwell M, Lisch D (2000). Transposable elements and host genome evolution. Trends in Ecology & Evolution 15: 95–99.

Kidwell MG (1983). Evolution of hybrid dysgenesis determinants in Drosophila melanogaster. Proceedings of the National Academy of Sciences 80: 1655–1659.

Kidwell MG, Kidwell JF, Sved JA (1977). Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. Genetics 86: 813–833.

Kim D, Langmead B, Salzberg SL (2015). HISAT: a fast spliced aligner with low memory requirements. Nat Meth 12: 357–360.

Kim N, Jinks-Robertson S (2012). Transcription as a source of genome instability. Nat Rev Genet 13: 204–214.

Kirkpatrick M, Hall DW (2004). Male-biased mutation, sex linkage, and the rate of adaptive evolution. Evolution 58: 437–440.

Kiuchi T, Koga H, Kawamoto M, Shoji K, Sakai H, Arai Y, et al. (2014). A single female-specific piRNA is the primary determiner of sex in the silkworm. Nature 509: 633–636.

Kozak KM, Wahlberg N, Neild AFE, Dasmahapatra KK, Mallet J, Jiggins CD

(2015). Multilocus species trees show the recent adaptive radiation of the mimetic Heliconius butterflies. Syst Biol 64: 505–524.

Kronforst MR, Hansen MEB, Crawford NG, Gallant JR, Zhang W, Kulathinal RJ, et al. (2013). Hybridization reveals the evolving genomic architecture of speciation. Cell Rep 5: 666–677.

Kruuk LE, Baird SJ, Gale KS, Barton NH (1999). A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. Genetics 153: 1959–1971.

Kuroda MI, Hilfiker A, Lucchesi JC (2016). Dosage Compensation in Drosophila - a Model for the Coordinate Regulation of Transcription. Genetics 204: 435–450.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. (2004). Versatile and open software for comparing large genomes. Genome Biol 5: R12.

Lafon-Placette C, Köhler C (2015). Epigenetic mechanisms of postzygotic reproductive isolation in plants. Curr Opin Plant Biol 23: 39–44.

Landry CR, Hartl DL, Ranz JM (2007). Genome clashes in hybrids: insights from gene expression. Heredity 99: 483–493.

Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL (2005). Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. Genetics 171: 1813–1822.

Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth 9: 357–359.

Laporte V, Charlesworth B (2002). Effective population size and population subdivision in demographically structured populations. Genetics 162: 501–519.

Lavoie CA, Platt RN, Novick PA, Counterman BA, Ray DA (2013). Transposable element evolution in Heliconius suggests genome diversity within Lepidoptera. Mob DNA 4: 21.

Layer RM, Chiang C, Quinlan AR, Hall IM (2014). LUMPY: a probabilistic framework for structural variant discovery. 15: R84.

Lenski RE, Ofria C, Pennock RT, Adami C (2003). The evolutionary origin of complex features. Nature 423: 139–144.

Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C (2017). TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. Nucleic Acids Research 45: e17.

Lewis SH, Quarles KA, Yang Y, Tanguy M, Frézal L, Smith SA, Sharma PP, Cordaux R, Gilbert C, Giraud I, Collins DH, Zamore PD, Miska EA, Sarkies P, Jiggins FM (2017) Pan-arthropod analysis reveals somatic piRNAs as an ancestral TE defence. Nature. In review.

Lexer C, Widmer A (2008). Review. The genic view of plant speciation: recent progress and emerging questions. Philos Trans R Soc Lond, B, Biol Sci 363: 3023–3036.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/pdf/1303.3997.pdf

Lima TG (2014). Higher levels of sex chromosome heteromorphism are associated with markedly stronger reproductive isolation. Nature

Communications 5: 4743.

Lin K, Smit S, Bonnema G, Sanchez-Perez G, de Ridder D (2015). Making the difference: integrating structural variation detection tools. Brief Bioinformatics 16: 852–864.

Long Y, Zhao L, Niu B, Su J, Wu H, Chen Y, et al. (2008). Hybrid male sterility in rice controlled by interaction between divergent alleles of two adjacent genes. Proc Natl Acad Sci USA 105: 18871–18876.

Lopez-Maestre H, Carnelossi EAG, Lacroix V, Burlet N, Mugat B, Chambeyron S, et al. (2017). Identification of misexpressed genetic elements in hybrids between Drosophila-related species. Sci Rep 7: 40618.

Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 15: 550.

Lucchesi JC, Kelly WG, Panning B (2005). Chromatin remodeling in dosage compensation. Annu Rev Genet 39: 615–651.

Luiz G (2008). Morphological Caste Studies In The Neotropical Swarm-Founding Polistinae Wasp. : 1–11.

Lunter G, Goodson M (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Research 21: 936–939.

Lynch M (2007). The evolution of genetic networks by non-adaptive processes. Nat Rev Genet 8: 803–813.

Lynch M (2010). Evolution of the mutation rate. Trends in Genetics 26: 345–352.

Lynch M, Conery JS (2000). The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.

Lynch M, Conery JS (2003). The origins of genome complexity. Science 302: 1401–1404.

Lynch M, Force A (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics 154: 459–473.

Lynch M, Milligan BG (1994). Analysis of population genetic structure with RAPD markers. Molecular Ecology 3: 91–99.

Lynch, M. 2007. The Origins of Genome Architecture. Sinauer Assocs., Inc., Sunderland, MA.

Mavarez, Audet, Bernatchez (2009). Major disruption of gene expression in hybrids between young sympatric anadromous and resident populations of brook charr (Salvelinus fontinalis). Journal of Evolutionary Biology 22: 1708–1720.

Mavarez J, Salazar CA, Bermingham E, Salcedo C, Jiggins CD, Linares M (2006). Speciation by hybridization in Heliconius butterflies. Nature 441: 868–871.

Mackay TFC (2010). Mutations and quantitative genetic variation: lessons from Drosophila. Philosophical Transactions of the Royal Society B: Biological Sciences 365: 1229–1239.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. (2012). The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178.

Madhani HD (2013). The frustrated gene: origins of eukaryotic gene expression. Cell 155: 744–749.

Magic traits in speciation: ‘magic’ but not rare? (2011). Magic traits in speciation: ‘magic’ but not rare? Trends in Ecology & Evolution 26: 389–397.

Malinsky M, Simpson JT, Durbin R. (2016). trio-sga: facilitating de novo

assembly of highly heterozygous genomes with parent-child trios. bioRxiv.

Mallet J (1995). A species definition for the modern synthesis. Trends in Ecology & Evolution 10: 294–299.

Mallet J (2005). Hybridization as an invasion of the genome. Trends in Ecology & Evolution 20: 229–237.

Mallet J (2006). What does Drosophila genetics tell us about speciation? Trends in Ecology & Evolution 21: 386–393.

Mallet J (2007). Hybrid speciation. Nature 446: 279–283.

Mallet J, Barton NH (1989). Strong natural selection in a warning-colour hybrid zone. Evolution 43: 421–431.

Mallet J. (1998). Species concepts. In Calow, P. (ed.) Encyclopaedia of Ecology and Environmental Management. Blackwell Press. pp. 709-711.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20: 229–237.

Mank JE (2009). Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. Am Nat 173: 141–150.

Mank JE (2013). Sex chromosome dosage compensation: definitely not for everyone. Trends in Genetics 29: 677–683.

Mank JE, Ellegren H (2009). All dosage compensation is local: gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. Heredity 102: 312–320.

Mank JE, Hosken DJ, Wedell N (2011). Some inconvenient truths about sex chromosome dosage compensation and the potential role of sexual conflict. Evolution 65: 2133–2144.

Mank JE, Nam K, Ellegren H (2010). Faster-Z evolution is predominantly due

to genetic drift. Molecular Biology and Evolution 27: 661–670.

Mank JE, Vicoso B, Berlin S, Charlesworth B (2010). Effective population size and the Faster-X effect: empirical results and their interpretation. Evolution 64: 663–674.

Manzanares M, Wada H, Itasaki N, Trainor PA, Krumlauf R, Holland PW (2000). Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head. Nature 408: 854–857.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, et al. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Research 23: 1817–1828.

Martin SH, Möst M, Palmer WJ, Salazar C, McMillan WO, Jiggins FM, et al. (2016). Natural Selection and Genetic Diversity in the Butterfly Heliconius melpomene. Genetics 203: 525–541.

Masly JP, Jones CD, Noor MAF, Locke J, Orr HA (2006). Gene Transposition as a Cause of Hybrid Sterility in Drosophila. Science 313: 1448–1450.

Masly JP, Presgraves DC (2007). High-resolution genome-wide dissection of the two rules of speciation in Drosophila. (NH Barton, Ed.). PLoS Biol 5: e243.

Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y (2007). Odorant-binding proteins OBP57d and OBP57e affect taste perception and host-plant preference in Drosophila sechellia. (MAF Noor, Ed.). PLoS Biol 5: e118.

Mayr E (1942) Systematics and the Origin of Species. Columbia Univ. Press, New York.

McClintock B (1953). Induction of Instability at Selected Loci in Maize. Genetics 38: 579–599.

McDermott SR, Noor MAF (2010). The role of meiotic drive in hybrid male

sterility. Philos Trans R Soc Lond, B, Biol Sci 365: 1265–1272.

McDonald JH, Kreitman M (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ (2010). Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Research 20: 816–825.

McMillan WO, Jiggins CD, Mallet J (1997). What initiates speciation in passion-vine butterflies? Proceedings of the National Academy of Sciences 94: 8628–8633.

Meisel RP, Connallon T (2013). The faster-X effect: integrating theory and data. Trends in Genetics 29: 537–544.

Meisel RP, Malone JH, Clark AG (2012). Faster-X evolution of gene expression in Drosophila. (D Bachtrog, Ed.). PLoS Genet 8: e1003013.

Merrill RM, Chia A, Nadeau NJ (2014). Divergent warning patterns contribute to assortative mating between incipient Heliconius species. Ecol Evol 4: 911–917.

Merrill RM, Dasmahapatra KK, Davey JW, Dell'Aglio DD, Hanly JJ, Huber B, et al. (2015). The diversification of Heliconius butterflies: what have we learned in 150 years? Journal of Evolutionary Biology 28: 1417–1438.

Merrill RM, Naisbit RE, Mallet J, Jiggins CD (2013). Ecological and genetic factors influencing the transition between host-use strategies in sympatric Heliconius butterflies. Journal of Evolutionary Biology 26: 1959–1967.

Merrill RM, Van Schooten B, Scott JA, Jiggins CD (2011). Pervasive genetic associations between traits causing reproductive isolation in Heliconius butterflies. Proceedings of the Royal Society B: Biological Sciences 278: 511–518.

Merrill RM, Wallbank RWR, Bull V, Salazar PCA, Mallet J, Stevens M, et al.

(2012). Disruptive ecological selection on a mating cue. Proc Biol Sci 279: 4907–4913.

Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD (2016). PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic Acids Research 44: D336–42.

Michalak P (2009). Epigenetic, transposon and small RNA determinants of hybrid dysfunctions. Heredity 102: 45–50.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. (2011). Mapping copy number variation by population-scale genome sequencing. Nature 470: 59–65.

Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, et al. (2015). The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Research 43: D213–21.

Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T (1987). Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harb Symp Quant Biol 52: 863–867.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37: 997–1002.

Mugal CF, Wolf JBW, Kaj I (2014). Why time matters: codon evolution and the temporal dynamics of dN/dS. Molecular Biology and Evolution 31: 212–231.

Muller H (1940). Bearing the Drosophila work on systematics. In: Huxley J (ed). The New Systematics. Clarendon Press: Oxford.

Muller, H.J. 1940. Bearing of the Drosophila work on systematics. In: The New Systematics (J. S. Huxley, ed.), pp. 185–268. Clarendon Press, Oxford.

Muller, H.J. 1942. Isolating mechanisms, evolution and temperature. Biol.

Symp. 6: 71–125.

Muñoz AG, Salazar C, Castaño J, Jiggins CD, Linares M (2010). Multiple sources of reproductive isolation in a bimodal butterfly hybrid zone. Journal of Evolutionary Biology 23: 1312–1320.

Nadeau NJ, Pardo-Diaz C, Whibley A, Supple MA, Saenko SV, Wallbank RWR, et al. (2016). The gene cortex controls mimicry and crypsis in butterflies and moths. Nature 534: 106–110.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, et al. (2011). Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Philosophical Transactions of the Royal Society B: Biological Sciences 367: 343–353.

Naisbit RE, Jiggins CD, Linares M, Salazar C, Mallet J (2002). Hybrid sterility, Haldane's rule and speciation in Heliconius cydno and Heliconius melpomene. Genetics 161: 1517–1526.

Naisbit RE, Jiggins CD, Mallet J (2001). Disruptive sexual selection against hybrids contributes to speciation between Heliconius cydno and Heliconius melpomene. Proceedings of the Royal Society B: Biological Sciences 268: 1849–1854.

Nam K, Munch K, Hobolth A, Dutheil JY, Veeramah KR, Woerner AE, et al. (2015). Extreme selective sweeps independently targeted the X chromosomes of the great apes. Proc Natl Acad Sci USA 112: 6413–6418.

Nguyen DK, Disteche CM (2006). Dosage compensation of the active X chromosome in mammals. Nat Genet 38: 47–53.

Nguyen LP, Galtier N, Nabholz B (2015). Gene expression, chromosome heterogeneity and the fast-X effect in mammals. Biology Letters 11: 20150010–20150010.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, et al. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. (C Tyler-Smith, Ed.). PLoS Biol 3: e170.

Noor MA, Grams KL, Bertucci LA, Reiland J (2001). Chromosomal inversions and the reproductive isolation of species. Proceedings of the National Academy of Sciences 98: 12084–12088.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009). Divergent selection and heterogeneous genomic divergence. Molecular Ecology 18: 375–402.

Nosil P, Schluter D (2011). The genes underlying the process of speciation. Trends in Ecology & Evolution 26: 160–167.

O'Neill MB, Mortimer TD, Pepperell CS (2015). Diversity of Mycobacterium tuberculosis across Evolutionary Scales. (SM Fortune, Ed.). PLoS Pathog 11: e1005257.

Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM (2009). The evolution of RNAi as a defence against viruses and transposable elements. Philos Trans R Soc Lond, B, Biol Sci 364: 99–115.

Obbard DJ, Welch JJ, Kim K-W, Jiggins FM (2009). Quantifying adaptive evolution in the Drosophila immune system. (DJ Begun, Ed.). PLoS Genet 5: e1000698.

Ohno S. (1970) Evolution by Gene Duplication. Springer-Verlag: New York, NY, USA.

Oka HI (1953)The mechanisms of sterility in the intervarietal hybrids. Phylogenetic differentiation of cultivated rice. VI. (In Japanese with English summary). Japan J Breed. 2:217-224

Oka HI (1957) Genic analysis for the sterility of hybrids between distantly related varieties of cultivated rice. J Genet. 55:397-409

Oka HI (1974). Analysis of genes controlling F1 sterility in rice by the use of

isogenic lines. Genetics 77:52-534.

Olds LC, Sibley E (2003). Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Hum Mol Genet 12: 2333–2340.

Orr HA (1993). Haldane's rule has multiple genetic causes. Nature 361: 532–533.

Orr HA (1995). The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics 139: 1805–1813.

Orr HA (2010). The population genetics of beneficial mutations. Philos Trans R Soc Lond, B, Biol Sci 365: 1195–1201.

Orr HA, Betancourt AJ (2001). Haldane's sieve and adaptation from the standing genetic variation. Genetics 157: 875–884.

Orr HA, Masly JP, Presgraves DC (2004). Speciation genes. Current Opinion in Genetics & Development 14: 675–679.

Orr HA, Turelli M (2001). The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. Evolution 55: 1085–1094.

Ortiz-Barrientos D (2005). Evidence for a One-Allele Assortative Mating Locus. Science 310: 1467–1467.

Ortiz-Barrientos D, Counterman BA, Noor MAF (2007). Gene expression divergence and the origin of hybrid dysfunctions. Genetica 129: 71–81.

Osanai-Futahashi M, Suetsugu Y, Mita K, Fujiwara H (2008). Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, Bombyx mori. Insect Biochemistry and Molecular Biology 38: 1046–1057.

Otto TD, Dillon GP, Degrave WS, Berriman M (2011). RATT: Rapid Annotation Transfer Tool. Nucleic Acids Research 39: e57–e57.

Panhuis TM, Clark NL, Swanson WJ (2006). Rapid evolution of reproductive proteins in abalone and Drosophila. Philosophical Transactions of the Royal Society B: Biological Sciences 361: 261–268.

Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, Joron M, et al. (2012). Adaptive introgression across species boundaries in Heliconius butterflies. (M R Kronforst, Ed.). PLoS Genet 8: e1002752.

Pardue ML, DeBaryshe PG (2011). Retrotransposons that maintain chromosome ends. Proc Natl Acad Sci USA 108: 20317–20324.

Parisi M, Nuttall R, Edwards P, Minor J, Naiman D, Lü J, et al. (2004). A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogaster adults. 5: R40.

Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, et al. (2003). Paucity of genes on the Drosophila X chromosome showing male-biased expression. Science 299: 697–700.

Paten B, Diekhans M, Earl D, John JS, Ma J, Suh B, et al. (2011). Cactus graphs for genome comparisons. J Comput Biol 18: 469–481.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D (2011). Cactus: Algorithms for genome multiple sequence alignment. Genome Research 21: 1512–1528.

Paudel Y, Madsen O, Megens H-J, Frantz LAF, Bosse M, Crooijmans RPMA, et al. (2015). Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. BMC Genomics 16: 330.

Phadnis N (2011). Genetic Architecture of Male Sterility and Segregation Distortion in Drosophila pseudoobscura Bogota-USA Hybrids. Genetics 189: 1001–1009.

Pinharanda A, Martin SH, Barker SL, Davey JW, Jiggins CD (2017). The comparative landscape of duplications in Heliconius melpomene and

Heliconius cydno. Heredity 118: 78–87.

Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Müller I, et al. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science 344: 1410–1414.

Pool JE, Nielsen R (2007). Population size changes reshape genomic patterns of diversity. Evolution 61: 3001–3006.

Presgraves DC (2002). Patterns of postzygotic isolation in Lepidoptera. Evolution 56: 1168–1183.

Presgraves DC (2010). The molecular evolutionary basis of species formation. Nat Rev Genet 11: 175–180.

Presgraves DC, Balagopalan L, Abmayr SM, Orr HA (2003). Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila. Nature 423: 715–719.

Presgraves DC, Stephan W (2007). Pervasive adaptive evolution among interactors of the Drosophila hybrid inviability gene, Nup96. Molecular Biology and Evolution 24: 306–314.

Price TD, Bouvier MM (2002). The evolution of F1 postzygotic incompatibilities in birds. Evolution 56: 2083–2089.

Pritchard JK, Di Rienzo A (2010). Adaptation - not by sweeps alone. Nat Rev Genet 11: 665–667.

Qian W, Zhang J (2014). Genomic evidence for adaptation by gene duplication. Genome Research 24: 1356–1362.

Quinlan AR, Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Ranwez V, Harispe S, Delsuc F, Douzery EJP (2011). MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop

codons. (WJ Murphy, Ed.). PLoS ONE 6: e22594.

Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL (2003). Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science 300: 1742–1745.

Ranz JM, Machado CA (2006). Uncovering evolutionary patterns of gene expression using microarrays. Trends in Ecology & Evolution 21: 29–37.

Rastogi A, Gupta D (2014). GFF-Ex: a genome feature extraction package. BMC Res Notes 7: 315.

Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28: i333–i339.

Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, et al. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. Journal of Evolutionary Biology 30: 1450–1477.

Reed RD, Papa R, Martin A, Hines HM, Counterman BA, Pardo-Diaz C, et al. (2011). optix drives the repeated convergent evolution of butterfly wing pattern mimicry. Science 333: 1137–1141.

Reiland J, Noor MAF (2002). Little qualitative RNA misexpression in sterile male F1 hybrids of Drosophila pseudoobscura and D. persimilis. BMC Evolutionary Biology 2: 16.

Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, et al. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nature Communications 4: 1–8.

Renaut S, Rowe HC, Ungerer MC, Rieseberg LH (2014). Genomics of homoploid hybrid speciation: diversity and transcriptional activity of long terminal repeat retrotransposons in hybrid sunflowers. Philos Trans R Soc Lond, B, Biol Sci 369: 20130345–20130345.

Rice P, Longden I, Bleasby A (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics 16: 276–277.

Rice WR (1984). Sex chromosomes and the evolution of sexual dimorphism. Evolution 38: 735–742.

Rigal M, Becker C, Pélissier T, Pogorelcnik R, Devos J, Ikeda Y, et al. (2016). Epigenome confrontation triggers immediate reprogramming of DNA methylation and transposon silencing in Arabidopsis thaliana F1 epihybrids. Proc Natl Acad Sci USA 113: E2083–92.

Rinn JL, Snyder M (2005). Sexual dimorphism in mammalian gene expression. Trends in Genetics 21: 298–305.

Rogers RL (2015). Tandem duplications and the limits of natural selection in Drosophila yakuba and Drosophila simulans. : 1–80.

Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR (2014). Landscape of Standing Variation for Tandem Duplications in Drosophila yakuba and Drosophila simulans. Molecular Biology and Evolution 31: 1750–1766.

Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR (2015). Tandem Duplications and the Limits of Natural Selection in Drosophila yakuba and Drosophila simulans. (A Palsson, Ed.). PLoS ONE 10: e0132184.

Romero-Soriano V, Modolo L, Lopez-Maestre H, Mugat B, Pessia E, Chambeyron S, et al. (2017). Transposable Element Misregulation Is Linked to the Divergence between Parental piRNA Pathways in Drosophila Hybrids. Genome Biology and Evolution 9: 1450–1470.

Rostant WG, Wedell N, Hosken DJ (2012). Transposable elements and insecticide resistance. Adv Genet 78: 169–201.

Rousselle M, Faivre N, Ballenghien M, Galtier N, Nabholz B (2016).

Hemizygosity Enhances Purifying Selection: Lack of Fast-Z Evolution in Two Satyrine Butterflies. Genome Biology and Evolution 8: 3108–3119.

Rundle HD, Nosil P (2005). Ecological speciation. Ecology Letters 8: 336–352.

Ryazansky S, Radion E, Mironova A, Akulenko N, Abramov Y, Morgunova V, et al. (2017). Natural variation of piRNA expression affects immunity to transposable elements. (C Feschotte, Ed.). PLoS Genet 13: e1006731.

Sackton TB, Corbett-Detig RB, Nagaraju J, Vaishna L, Arunkumar KP, Hartl DL (2014). Positive selection drives faster-Z evolution in silkmoths. Evolution 68: 2331–2342.

Schilthuizen M, Giesbers MCWG, Beukeboom LW (2011). Haldane's rule in the 21st century. Heredity 107: 95–102.

Seehausen O, BUTLIN RK, Keller I, Wagner CE, BOUGHMAN JW, Hohenlohe PA, et al. (2014). Genomics and the origin of species. Nat Rev Genet 15: 176–192.

Servedio MR, Noor MAF (2003). T HER OLE OFR EINFORCEMENT INS PECIATION: Theory and Data. Annu Rev Ecol Evol Syst 34: 339–364.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, et al. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428: 717–723.

Singh G, Popli S, Hari Y, Malhotra P, Mukherjee S, Bhatnagar RK (2009). Suppression of RNA silencing by Flock house virus B2 protein is mediated through its interaction with the PAZ domain of Dicer. FASEB J 23: 1845–1857.

Singh ND, Koerich LB, Carvalho AB, Clark AG (2014). Positive and purifying selection on the Drosophila Y chromosome. Molecular Biology and Evolution 31: 2612–2623.

Singh R, Jagadeeshan S (2012). Sex and speciation: Drosophila reproductive tract proteins- twenty five years later. International Journal of Evolutionary Biology 2012: 191495–9.

Singh R, Jagadeeshan S (2012). Sex and speciation: Drosophila reproductive tract proteins- twenty-five years later. International Journal of Evolutionary Biology 2012: 191495–9.

Sjödin P, Jakobsson M (2012). Population genetic nature of copy number variation. Methods Mol Biol 838: 209–223.

Slotkin RK, Martienssen R (2007). Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8: 272–285.

Smeds L, Warmuth V, Bolivar P, Uebbing S, Burri R, Suh A, et al. (2015). Evolutionary analysis of the female-specific avian W chromosome. Nature Communications 6: 7330.

Smit AFA, Hubley R, Green P. 2013. RepeatMasker Open-4.0. (2013–2015) Available from http://www.repeatmasker.org.

Susumu Ohno (1970). Evolution by gene duplication. Springer-Verlag. ISBN 0-04-575015-7.

Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD (2012). A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. Nat Protoc 7: 1260–1284.

Sweigart AL, Flagel LE (2015). Evidence of natural selection acting on a polymorphic hybrid incompatibility locus in Mimulus. Genetics 199: 543–554.

Swevers L, Iatrou K (2003). The ecdysone regulatory cascade and ovarian development in lepidopteran insects: insights from the silkmoth paradigm. Insect Biochemistry and Molecular Biology 33: 1285–1297.

Sánchez-Gracia A, Maside X, Charlesworth B (2005). High rate of horizontal

transfer of transposable elements in Drosophila. Trends in Genetics 21: 200–203.

Tang S, Presgraves DC (2009). Evolution of the Drosophila nuclear pore complex results in multiple hybrid incompatibilities. Science 323: 779–782.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics 31: 2032–2034.

Tattini L, D'Aurizio R, Magi A (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front Bioeng Biotechnol 3: 92.

Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics 28: 2711–2718.

The Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature: 1–5.

Thornton K, Bachtrog D, Andolfatto P (2006). X chromosomes and autosomes evolve at similar rates in Drosophila: no evidence for faster-X protein evolution. Genome Research 16: 498–504.

Thornton K, Long M (2002). Rapid divergence of gene duplicates on the Drosophila melanogaster X chromosome. Molecular Biology and Evolution 19: 918–925.

Ting C-T, Tsaur S-C, Sun S, Browne WE, Chen Y-C, Patel NH, et al. (2004). Gene duplication and speciation in Drosophila: evidence from the Odysseus locus. Proceedings of the National Academy of Sciences 101: 12232–12235.

Turelli M, Begun DJ (1997). Haldane's rule and X-chromosome size in Drosophila. Genetics 147: 1799–1815.

Turelli M, Moyle LC (2006). Asymmetric Postmating Isolation: Darwin‘s

Corollary to Haldane's Rule. Genetics 176: 1059–1088.

Turelli M, Orr HA (2000). Dominance, epistasis and the genetics of postzygotic isolation. Genetics 154: 1663–1679.

Turner TL, Hahn MW, Nuzhdin SV (2005). Genomic islands of speciation in Anopheles gambiae. (N Barton, Ed.). PLoS Biol 3: e285.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. (2005). Fine-scale structural variation of the human genome. Nat Genet 37: 727–732.

Úbeda F, Patten MM, Wild G (2015). On the origin of sex chromosomes from meiotic drive. Proc Biol Sci 282: 20141932–20141932.

Van Belleghem SM, Rastas P, Papanicolaou A, Martin SH, Arias CF, Supple MA, et al. (2017). Complex modular architecture around a simple toolkit of wing pattern genes. Nat Ecol Evol 1: 0052.

Vicoso B, Charlesworth B (2006). Evolution on the X chromosome: unusual patterns and processes. Nat Rev Genet 7: 645–653.

Vicoso B, Charlesworth B (2009). Effective population size and the faster-X effect: an extended model. Evolution 63: 2413–2426.

Vicoso B, Kaiser VB, Bachtrog D (2013). Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. Proc Natl Acad Sci USA 110: 6453–6458.

Wagner A (2000). Robustness against mutations in genetic networks of yeast. Nat Genet 24: 355–361.

Wallbank RWR, Baxter SW, Pardo-Diaz C, Hanly JJ, Martin SH, Mallet J, et al. (2016). Evolutionary Novelty in a Butterfly Wing Pattern through Enhancer Shuffling. (NH Barton, Ed.). PLoS Biol 14: e1002353.

Walters JR, Hardcastle TJ, Jiggins CD (2015). Sex Chromosome Dosage

Compensation in Heliconius Butterflies: Global yet Still Incomplete? Genome Biology and Evolution 7: 2545–2559.

Wang X, Weigel D, Smith LM (2013). Transposon variants and their effects on gene expression in Arabidopsis. (GP Copenhaver, Ed.). PLoS Genet 9: e1003255.

White, M. J. D. 1978. Modes of speciation. W. H. Freemen. San Francisco.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. (2007). A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973–982.

Wilson, EO. (1992) The Diversity of Life. Belknap.

Witherspoon DJ, Watkins WS, Zhang Y, Xing J, Tolpinrud WL, Hedges DJ, et al. (2009). Alu repeats increase local recombination rates. BMC Genomics 10: 530.

Wittkopp PJ, Haerum BK, Clark AG (2004). Evolutionary changes in cis and trans gene regulation. Nature 430: 85–88.

Wittkopp PJ, Haerum BK, Clark AG (2008). Regulatory changes underlying expression differences within and between Drosophila species. Nature Publishing Group 40: 346–350.

Wolf JBW, Ellegren H (2017). Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet 18: 87–100.

Wolf JBW, Lindell J, Backström N (2010). Speciation genetics: current status and evolving approaches. Philos Trans R Soc Lond, B, Biol Sci 365: 1717–1733.

Wolfe KH (2001). Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2: 333–341.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, et al.

(2003). The evolution of transcriptional regulation in eukaryotes. Molecular Biology and Evolution 20: 1377–1419.

Wright AE, Zimmer F, Harrison PW, Mank JE (2015). Conservation of Regional Variation in Sex-Specific Sex Chromosome Regulation. Genetics: genetics.115.179234.

Yao W. (2015). intansv: Integrative analysis of structural variations. R package version 1.9.2.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865–2871.

Zhang L, Lu HHS, Chung W-Y, Yang J, Li W-H (2005). Patterns of segmental duplication in the human genome. Molecular Biology and Evolution 22: 135–141.

Zhang Z, Hambuch TM, Parsch J (2004). Molecular evolution of sex-biased genes in Drosophila. Molecular Biology and Evolution 21: 2130–2139.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 14 Suppl 11: S1.

Zhou Q, Bachtrog D (2012). Sex-specific adaptation drives early sex chromosome evolution in Drosophila. Science 337: 341–345.

Zichner T, Garfield DA, Rausch T, Stutz AM, Cannavo E, Braun M, et al. (2013). Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. Genome Research 23: 568–579.

Appendix A

The comparative landscape of duplication in *Heliconius melpomene* and *Heliconius cydno*

Heredity (2017) 118, 78–87
Official journal of the Genetics Society
www.nature.com/hdy

OPEN

ORIGINAL ARTICLE

The comparative landscape of duplications in *Heliconius melpomene* and *Heliconius cydno*

A Pinharanda, SH Martin, SL Barker, JW Davey and CD Jiggins

Gene duplications can facilitate adaptation and may lead to interpopulation divergence, causing reproductive isolation. We used whole-genome resequencing data from 34 butterflies to detect duplications in two *Heliconius* species, *Heliconius cydno* and *Heliconius melpomene*. Taking advantage of three distinctive signals of duplication in short-read sequencing data, we identified 744 duplicated loci in *H. cydno* and *H. melpomene* and evaluated the accuracy of our approach using single-molecule sequencing. We have found that duplications overlap genes significantly less than expected at random in *H. melpomene*, consistent with the action of background selection against duplicates in functional regions of the genome. Duplicate loci that are highly differentiated between *H. melpomene* and *H. cydno* map to four different chromosomes. Four duplications were identified with a strong signal of divergent selection, including an odorant binding protein and another in close proximity with a known wing colour pattern locus that differs between the two species.
Heredity (2017) **118,** 78–87; doi:10.1038/hdy.2016.107; published online 7 December 2016

INTRODUCTION

Gene duplications occur frequently in eukaryotic genomes, where duplication rates are on the order of 0.01 per gene per million years (Lynch and Conery, 2000). Duplication is considered to be the main mechanism by which new genes arise (Katju, 2012), providing material for the origin of evolutionary novelties (Hunt *et al.*, 1998; Manzanares *et al.*, 2000; Kassahn *et al.*, 2009; Qian and Zhang, 2014). For example, the frequency of gene copy-number variants (CNVs) increased during experimental evolution experiments in *Caenorhabditis elegans* (Farslow *et al.*, 2015) and, in *Escherichia coli*, a tandem gene duplication was responsible for the evolutionary novelty in citrate metabolism seen in the long-term evolution experiment (Blount *et al.*, 2012). Such variation shapes gene expression profiles and influences phenotypic diversity (Feuk *et al.*, 2006; Iskow *et al.*, 2012; Katju and Bergthorsson, 2013).

The most common outcome for gene duplicates is to become pseudogenes through the accumulation of deleterious mutations (Lynch and Conery, 2000). Preservation of duplicate genes by natural selection may depend on whether or not one of the two gene copies accumulates mutations that lead to novel beneficial functions (Ohno, 1970). For example, trichromatic vision in Old World primates evolved by duplication of an X-linked opsin gene, an example of *neofunctionalization* (Hunt *et al.*, 1998). In addition, preservation of gene duplicates by natural selection may also occur by selection for increasing gene dosage as shown for ancient duplicates of *Saccharomyces cerevisiae* (Conant and Wolfe, 2008) or for regulatory robustness (Keane *et al.*, 2014). The duplication event does not, however, need to span the complete length of the gene. For example, a partial gene duplication is responsible for the origin of the antifreeze glycoprotein in Antarctic fish (Deng *et al.*, 2010). Alternatively, in *subfunctionalization* models, duplicates are preserved through each copy adopting a subset of the functions of the ancestral gene (Lynch and Force, 2000). This might occur when, for example, regulatory elements of the duplicate loci accumulate mutations that enable both duplicates to take on new functions different to that of the ancestral gene. In zebrafish, *engrailed-1* and *-1b* are a duplicate pair of transcription factors that evolved complementary expression patterns (Force *et al.*, 1999).

Gene duplication can also contribute to speciation. Duplicate genes can provide the raw material for populations to evolve divergent strategies and adapt to novel habitats, or may lead to genetic incompatibilities (Ting *et al.*, 2004). As such, diversification in gene function between duplicated genes can potentially contribute to reproductive isolation. In *Arabidopsis thaliana* recessive embryo lethality is explained by the divergent evolution of two paralogues of a duplicate gene important for the catalyses of the biosynthetic pathway producing histidine. The reciprocal gene loss has led to genetic incompatibilities in specific crosses (Bikard *et al.*, 2009).

Historically, CNVs were identified with cytogenetic technologies such as fluorescence *in situ* hybridization and karyotyping. More recently, array-based comparative genomic hybridization and single-nucleotide polymorphism array approaches have been used. However, array experiments have several weaknesses including limited coverage of the genome, hybridization noise and difficulty in detecting novel and rare variants (Zhao *et al.*, 2013). It is now possible to detect CNVs using next-generation sequencing technology that generates millions of randomly sampled short (100–300 bp) reads in a single run. Several methods have been developed to detect CNVs from short-read data: (1) analysis of abnormally mapping read pairs (paired-end (PE)); (2) analysis of the number of reads aligned to regions of the genome, or read depth (RD); (3) analysis of clipped/gapped alignments, or split reads (SRs); and (4) *de novo* assembly of resequenced genomes

Department of Zoology, University of Cambridge, Cambridge, UK
Correspondence: A Pinharanda or Professor CD Jiggins, Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK.
E-mail: alp66@cam.ac.uk or c.jiggins@zoo.cam.ac.uk
Received 16 March 2016; revised 24 August 2016; accepted 25 August 2016; published online 7 December 2016

(Ye *et al.*, 2009; Abyzov *et al.*, 2011; Rausch *et al.*, 2012; Chen *et al.*, 2014). In order to increase the accuracy and confidence of the calls, a common approach is to integrate the different strategies into a pipeline where complementary signals are incorporated (Mills *et al.*, 2011; Lin *et al.*, 2015; Tattini *et al.*, 2015; Teo *et al.*, 2012). CNVs have now been surveyed across the genomes of a range of closely related species or populations such as sticklebacks, pea-aphids, pigs and fruit-flies (Chain *et al*, 2014; Feulner *et al.*, 2013; Duvaux *et al.*, 2015; Paudel *et al.*, 2015; Rogers *et al.*, 2015).

Here we investigate duplications in the genomes of two species of Neotropical *Heliconius* butterfly. This taxonomic group has been studied for over 150 years since the first evolutionists became fascinated with their striking wing pattern diversity. Since then, *Heliconius* has contributed to answering evolutionary questions covering a broad range of research topics from taxonomy to ecology, behaviour and genetics (Merrill *et al.*, 2015). The best studied species pair are *Heliconius cydno* and *Heliconius melpomene*, two hybridizing sympatric species that differ in their ecology, mimicry patterns and mate preferences. They show low levels of inter-specific hybridization that nonetheless results in genome-wide signatures of admixture (Martin *et al.*, 2013). An outstanding question remains over the number and identity of the genomic regions that contribute to their speciation.

Genetic studies of *Heliconius* butterflies have focussed on loci controlling colour patterns, with many races diverging at these loci alone (Nadeau *et al.*, 2011; Martin *et al.*, 2013). Strong and rapid ecological divergence seems to be a driver of the earliest stages of speciation (Jiggins *et al.*, 2001; McMillan *et al.*, 1997; Muñoz *et al.*, 2010). However, recently, gene duplication in the genus has been linked to the evolution of visual complexity, development and immunity (The Heliconius Genome Consortium, 2012), as well as female oviposition behaviour (Briscoe *et al.*, 2013). Moreover, Nadeau *et al.* (2011) identified multiple CNVs between different *Heliconius* races. These results make *Heliconius* butterflies a promising system for an investigation of evolution by gene duplication for both autosomal and sex-linked genes.

We identify duplications using PE, SR and RD information from whole-genome resequencing short-read data for two *Heliconius* species, *H. cydno* and *H. melpomene*, using a similar strategy to the one used to discover and genotype structural variants in the human 1000 Genomes Project (Mills *et al.*, 2011) and the *Drosophila melanogaster* Genetic Reference Panel (Zichner *et al.*, 2013). By integrating different variant calling algorithms, and taking advantage of three distinctive next-generation sequencing signals, we map duplications among wild-caught *Heliconius* samples from two different species and three different locations, and identify loci putatively under divergent selection that may play a role in speciation.

MATERIALS AND METHODS

DNA sequence data retrieval and mapping of short-read data

Illumina (San Diego, CA, USA) paired-end sequencing data for 20 *H. melpomene* and 14 *H. cydno* butterflies (SRA106228, Kronforst *et al.*, 2013; ERP002440, Martin *et al.*, 2013) was downloaded from public repositories using the NCBI SRA toolkit (v2.5.7; National Center for Biotechnology Information, Bethesda, MD, USA). The reads were aligned to the *H. melpomene* genome (v2.0) (Davey *et al.*, 2016) with Stampy (v1.0.23; Lunter and Goodson, 2011) using default values for all parameters except the substitution rate, which was set to 0.01. Picard (v1.128) (picard.sorceforge.net)) was used to convert SAM/BAM files and remove PCR duplicate read pairs. Bcftools (v1.3; Li *et al.*, 2009) and bedtools (v2.20.1-13-g9249816; Quinlan and Hall, 2010) were used to process BAM and VCF files (Supplementary Table S1).

Detecting duplications through the analysis of SR, PE and RD information

The structural variant discovery methods DELLY (v0.6.1) (Rausch *et al.*, 2012), CNVnator (v0.3.2) (Abyzov *et al.*, 2011) and Pindel (v0.2.5a7) (Ye *et al.*, 2009) were used to detect candidate duplications in a focal set of 10 *H*eliconius *melpomene rosina* and 10 *Heliconius cydno galanthus* from Costa Rica, representing the largest population sample available for each species. We ran DELLY and Pindel on each population and CNVnator on each sample individually. These algorithms analyse different sequence signals to call the putative duplications: DELLY uses SR and PE information, Pindel uses SR information and CNVnator uses RD variation. CNVnator was run with a bin size of 100 bp, as recommended by the authors of the software, and all other parameters were set to default values (Table 1, raw calls). For simplicity, we focus on duplications and do not report deletions in the resequenced individuals relative to the reference.

The three methods we used to generate our Discovery Sets (PE, RD and SRs) required mapping to a reference genome. Duplication of loci in the reference genome has been shown to influence the discovery of structural variants and the alignment strategy used is important in detecting duplications in repeated regions (Teo *et al.*, 2012). There were several different alignment strategies we could have chosen to deal with reads mapping to more than one location. It was possible to (1) discard these reads, (2) report all possible positions to which the reads map and (3) choose a position at random out of all equally good matching positions.

Limiting the analysis to uniquely mapped regions of the genome (strategy 1) would be likely to miss duplications, especially considering the high heterozygosity of these samples. Using algorithms that consider all possible mapping locations (strategy 2) has not been tested in samples where the mean RD is

Table 1 Duplication discovery and genotyping in *Heliconius cydno* and *Heliconius melpomene*

Species	*Method*	*Raw calls*	*Merged by tool*	*Discovery set: merged by species*	*Genotyping set*	*Heliconius set*
H. cydno	DELLY (PE and SR)	14 691	5883			
	CNVnator (RD)	20 936	6376	1920	497	
	Pindel (SR)	1 261 451	15 611			744
H. melpomene	DELLY (PE and SR)	21 870	5097			
	CNVnator (RD)	22 267	10 751	1591	463	
	Pindel (SR)	896 202	7889			

Abbreviations: PE, paired-end; RD, read depth; SR, split read.
Duplication discovery sets were generated by mapping duplications in *H. cydno* and *H. melpomene* using whole-genome re-sequencing data from 20 wild Costa-Rican individuals (10 *H. cydno galanthus* and 10 *H. melpomene rosina*) (Discovery Set). A further 14 wild individuals from Panama (4 *H. cydno chioneus*, 4 *H. melpomene rosina* and 6 *H. melpomene melpomene*) were used to generate each of the species-specific genotyping sets (Genotyping Set). Both genotyping sets were merged and any resulting redundant calls filtered. This resulted in 744 duplications segregating in the Heliconius set.

lower than 20× (Teo *et al.*, 2012). All the samples we used to generate our Discovery Sets were sequence to an average of 15× and hence we chose not to use this strategy. Placing a read at random when all the possible positions are an equally good match (strategy 3) has been shown to dilute the signal of duplications (Teo *et al.*, 2012). However, because this strategy has been used extensively in previous work and is a conservative strategy, we chose this over the other approaches (Zichner *et al.*, 2013).

Filtering and merging duplication predictions: the discovery sets

To generate a list of non-redundant duplications for each species we combined the predictions generated by the three methods using custom scripts (available from Dryad) (Figure 1a). We calculated confidence intervals around each putative breakpoint according to the resolution defined for each method (DELLY: 50 bp outwards, 100 bp inwards; CNVnator: 1 kb outwards, 400 bp inwards; Pindel: +/− 10 bp) (Zichner *et al.*, 2013) (Table 1, merged by tool; Figure 1a). We generated six duplication discovery call sets (one for each combination of three methods and two species) by combining all calls with overlapping confidence intervals at both start and end coordinates into a single event. Predictions made by DELLY had to have at least three read-pairs with a mapping quality higher than 20 supporting the call for each individual sample. We removed 311 duplication calls that were predicted by DELLY in all of the *H. melpomene* samples, and were therefore likely to represent either genome assembly errors or genuine deletions in the reference genome. Finally, we combined the three putative call sets within each species using the intansv module (v1.9.2) in R (v3.2.1; https://cran.r-project.org; Yao, 2015). We kept calls that had a reciprocal coordinate overlap of 90% or higher and were predicted by at least two methods. Previous studies had used an overlap of 80% (Zichner *et al.*, 2013). However, because the size and total count of the putative variants did not differ dramatically between cut offs of 80 and 90% in our data set (Supplementary Figures S1–S4), we chose to use 90% as a more conservative overlap parameter. This generated two species-specific duplication discovery call sets, one for *H. cydno* and one for *H. melpomene* (Table 1, Discovery Set; Figure 1a, Discovery Sets).

Duplication genotype calling: the genotyping sets

To infer copy-number genotypes and evaluate the occurrence of each duplication in both Discovery Sets for all samples (20 *H. melpomene* and 14 *H. cydno*), we used the DELLY genotyper module with –t DUP option and default parameters (v0.7.2) (Rausch *et al.*, 2012). All duplications were treated as dominant loci and genotypes were scored as presence or absence in each sample. Using svprops, a program that computes various SV statistics from an input vcf file (https://github.com/tobiasrausch/svprops), we calculated median read support of each variant. We filtered out duplications with more than 500 reads mapping in an effort to discard repeats found at high copy number throughout the genome. We also filtered out events not genotyped in any of the samples, leaving high-quality Genotyping Sets of 497 putative duplications in *H. cydno* and 462 in *H. melpomene* (Figure 1a, Genotyping Sets).

Merging the *H. melpomene* and *H. cydno* genotyping sets: the Heliconius set

There were 186 identified putative duplications in the Genotyping Set of *H. melpomene* and *H. cydno* with an overlap >90% and these were merged further using the intansv module (v1.9.2) in R (v3.2.1) (Yao, 2015). After merging both Genotyping Sets according to this criterion we produced the Heliconius Set (Figure 1). Each duplication event was treated as a dominant binary marker (0 for absence and 1 for presence). A duplication was considered to be absent (0) when individual *i* has the same number of copies of sequence *j* as the Hmel2 reference genome, whatever the number of *j* copies in the reference genome. Conversely, a duplication was considered to be present (1) when *i* has more copies of *j* than the Hmel2 reference genome. We called genotypes as presence/absence in this way, rather than calling heterozygotes (Rausch *et al.*, 2012).

Inferring the quality of the putative calls by PacBio alignment and analysis of chromosome 2

We evaluated the accuracy of our duplication calling methods on a separate set of individuals for which appropriate long-read sequence data were available. These were one *H. melpomene* and one *H. cydno* family, for which the parents and one offspring from each family had been sequenced on an Illumina HiSeq 2000 (125 bp paired end, ENA accession ERP009507; see Malinsky *et al.*, 2016 for details). Our full duplication detection pipeline was run on these six individuals for chromosome 2. In addition, pools of 12 female and 12 male larvae from the same two families were sequenced on a Pacific Biosciences (PacBio, Menlo Park, CA, USA) RS II machine (P6/C4 chemistry, ENA submission in progress; read depths: *H. melpomene* females, 54x; *H. melpomene* males, 37x; *H. cydno* females, 49x; *H. cydno* males, 14x). Pacific Biosciences sequences were aligned to the *H. melpomene* reference genome version 2.0 (Davey *et al.*, 2016) with bwa mem (Li, 2013), using the PacBio option (-x). We then followed Layer *et al.* (2014) to validate our putative duplications, using sambamba (v0.6.1, Tarasov *et al.*, 2015) to select and filter the SRs from each PacBio bam file and converting these to the bedpe format (v2.25.0) (Quinlan and Hall, 2010) using the LUMPY (https://github.com/arq5x/lumpy-sv) custom script splitReadSamToBedpe. To convert the SRs to breakpoint calls we ran the custom script splitterToBreakpoint on each bedpe file with slope 1000 and default options for all other parameters (Layer *et al.*, 2014). The bedpe files with breakpoint information were merged for each species using bedtools intersectBed (v2.25.0) (Quinlan and Hall, 2010). We selected those reads that overlapped the start and end of the putative breakpoints called using Illumina short-read data. A putative duplication was considered validated when there were split long-read alignments within the predicted breakpoint interval such that (1) two segments of a single PacBio subread aligned to overlapping sections of the reference (Figure 2, PacBio read R1); or (2) if a single read aligned in split formation with the downstream end of the read aligning to a region that is upstream in the reference (Figure 2, PacBio read R2) (Layer *et al.*, 2014; Rogers *et al.*, 2014).

Using the putative genotyping duplication call set to show population structure and differentiation

Putative duplications from the Heliconius Set were analysed as dominant loci by principal component analysis in using the R package adegenet (v1.3-1) (Figure 3; Armengol *et al.*, 2009; Jombart and Ahmed, 2011).

Overlap between structural variants and genomic features

We investigated the overlap between the genotyped duplications and four different genomic features (genes, coding sequences (CDSs), introns and untranslated regions (UTRs)) using the R package 'intervals' in both Genotyping sets (Figure 1a and Table 1 Genotyping set). A single duplication could fall into several subcategories. All duplications that overlapped with coding sequence were counted as CDS duplications. A duplication was considered to be intronic if it overlapped with an intron but not CDS. UTRs were considered in the same way as introns if it does not overlap with CDS. Overlap with any of these features was considered a gene-overlapping duplication. As a small number of the genotyped duplications were overlapping, these were merged for this analysis, so that only non-overlapping duplication intervals were considered. To investigate whether the observed number of duplications overlapping each class of genomic features was significantly larger or smaller than expected by chance, we simulated 10 000 randomized distributions of duplications across the genome. In each simulation, the defined set of duplication intervals (with overlapping intervals merged for simplicity) was randomly permuted into non-overlapping locations across the genome, and the number overlapping with each class of genomic feature was recorded. We used the 2.5 and 97.5% quantiles of the simulated distribution as critical values to assess whether the observed overlaps differed significantly from that expected under a random distribution of duplications.

Detection of enriched biological functions within the Heliconius Set

We used InterProScan (v5.18.57.0; https://www.ebi.ac.uk/interpro/) (options –t n –goterms) to compare the Heliconius Set against the InterPro database. The InterPro database integrates predictive information from a number of sources (Mitchell *et al.*, 2015). We analysed PANTHER (http://www.pantherdb.org)

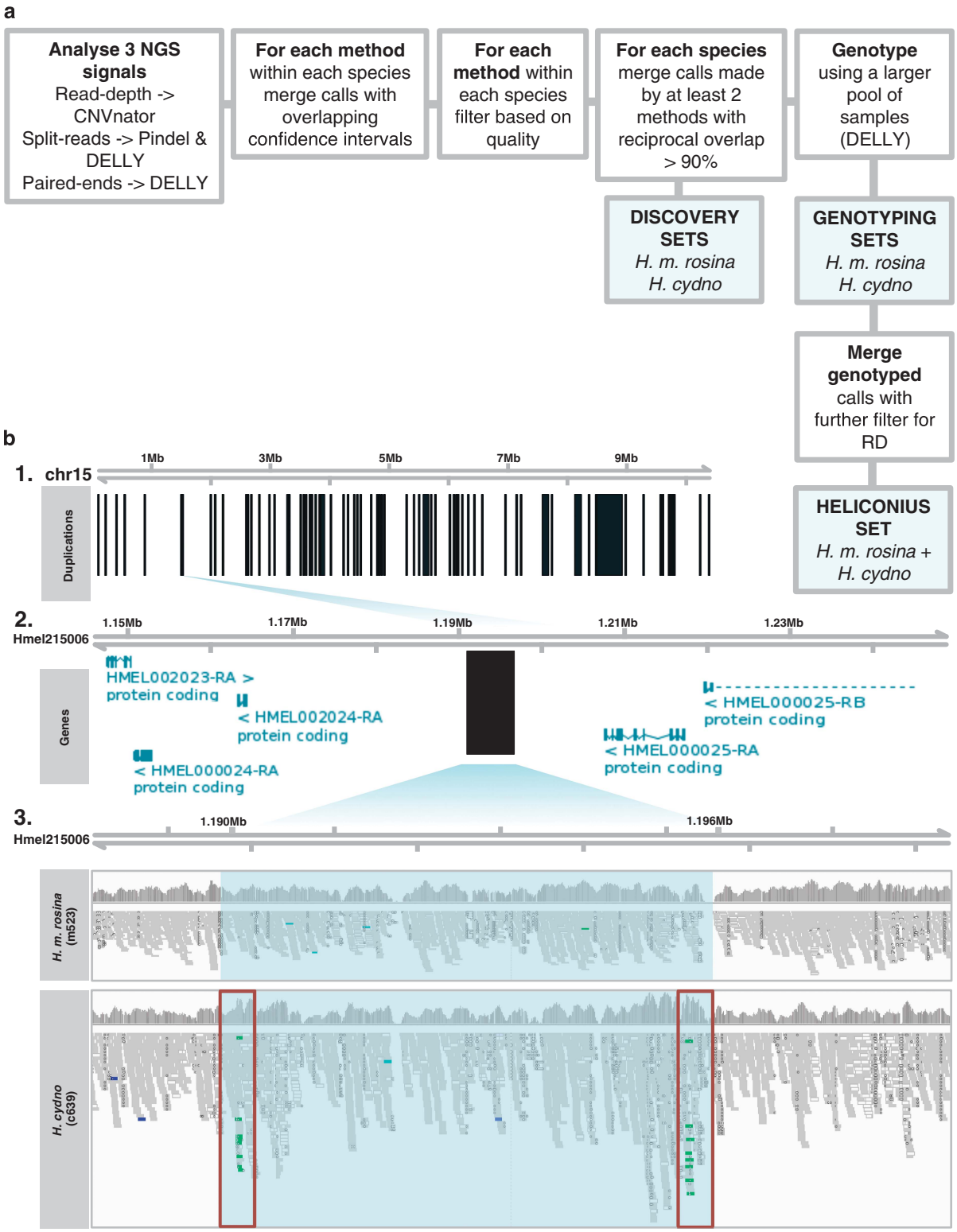


Figure 1 Duplication mapping and genotyping. (**a**) Integrated pipeline for duplication discovery (Discovery Sets) and genotyping (Genotyping Sets). Heliconius Set is the merged and filtered Genotyping sets from *H. cydno* and *H. melpomene*. (**b**) Example of a polymorphic duplication in *H. cydno* with respect to the *H. m. melpomene* reference genome (Davey *et al.*, 2016). (**b**1) Schematic representation of merged and genotyped Heliconius set duplication (vertical black rectangles) in Heliconius set for chromosome 15 (Table 1, Heliconius set). (**b**2) Zoom-in scaffold Hmel215006 to focus on a putative duplication from the merged genotyped set mapping 5′ end of the gene *cortex* (Nadeau *et al.*, 2016) (Table 3, Hmel215006:1190144-1196212). HMEL000025-RA and HMEL000025-RB are transcripts of *cortex* that map to Hmel215006:1205164-1324501. Genes flanking the duplication annotated as in Hmel2 (Davey *et al.*, 2016). (**b**3) Zooming-in further and looking at IGV RD and Illumina tracks for one *H. melpomene* and one *H. cydno* sample. Shaded light-blue region delineates the region that was identified as being duplicated. Red rectangles correspond to the breakpoint location of the region. Tracks are coloured green when a tandem duplication with respect to the reference genome is predicted by the read-pair orientation (PE) information.

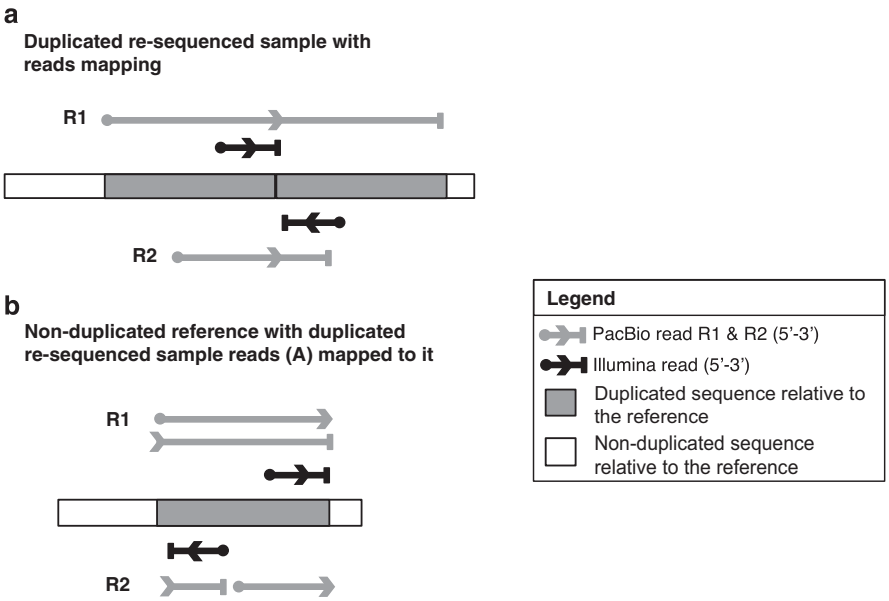
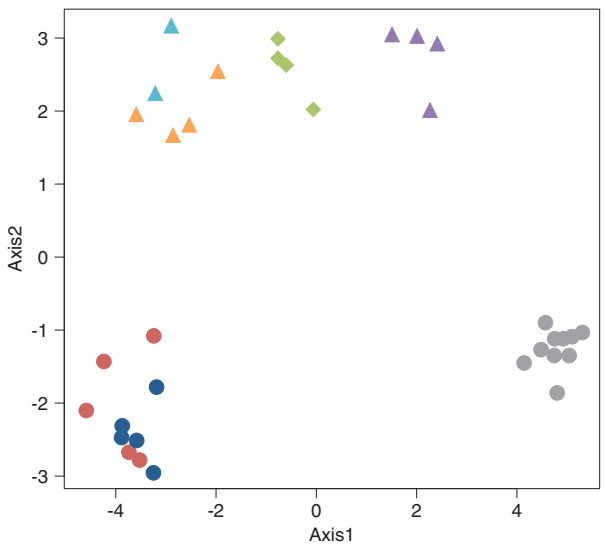


Figure 2 Validating short-read calls on chromosome 2 using PacBio single-molecule sequencing. Example of a breakpoint structure associated with a tandem duplication sequenced by Illumina chemistry (short reads, black) and PacBio chemistry (long reads, grey). A circle denotes the start of a read, the arrow its orientation, and the end is represented by a vertical bar. PacBio read R1 spans the entire duplicated sequence but PacBio read R2 does not. (**a**) Duplicated resequenced sample with Illumina and PacBio reads (R1 and R2) mapping. (**b**) Non-duplicated reference with duplicated resequenced sample reads from A mapped to it—tandem duplicated sequence aligned to a non-duplicated reference. Illumina reads from an individual with a tandem duplication map in divergent orientations when aligned to a reference without duplicated sequence. When PacBio read R1 is aligned to a non-duplicated reference, there are two alignments to the region that is flanked by the Illumina divergently oriented reads. The PacBio read R2 aligns discontinuously to the reference genome. The 3′ end of the R2 fragment of the breakpoint aligns to the reference upstream of the 5′ end of the R2 fragment.



Species	Location
H. m. melpomene	French Guiana
H. m. melpomene	Panama
H. m. rosina	Panama
H. cydno chioneus	Panama
H. cydno galanthus	Costa Rica
H. m. rosina	Costa Rica (Pacific)
H. m. rosina	Costa Rica (Caribbean)

Figure 3 Principal component analysis of the duplicated variants in the Heliconius set. Samples cluster by species and location based on their duplication genotype. Of the total variance, 17.57% was explained by the first two principal components (PC1 12.97% and PC2 4.6%).

database IDs that can be used to infer the function of uncharacterized genes based on their evolutionary relationships to genes with known functions (Mi *et al.*, 2016). We ran the PANTHER overrepresentation test on the Heliconius Set using the *D. melanogaster* genome as the reference list. We performed this analysis on the PANTHER GO-Slim Biological Process. We used the Bonferroni correction for multiple testing and report those categories overrepresented with $P<0.05$ (Supplementary Table S2 and Supplementary Figure S13). Five hundred and twenty nine overrepresented occurrences did not have a biological process associated with them but we have reported their predicted family name (Supplementary Table S3).

Identifying outlier loci from the Heliconius Set

Duplications present in the Heliconius Set were tested for signals of divergent selection by identifying F_{ST} outliers using BayeScan (v2.1; Foll and Gaggiotti, 2008) with default parameters except that prior odds were set to 1 (Cheang *et al.*, 2013). F_{ST} was estimated for the Heliconius Set between (1) *H. cydno* Costa (Rica and Panama); and (2) *H. melpomene* (Costa Rica, Panama and French Guiana). Each duplication event was treated as a dominant binary marker (0 for absence and 1 for presence). We corrected for false positives (false discovery rate of $P<0.05$). Duplications with log posterior odds >1 have strong support for selection.

We also applied a related method that identifies loci subject to selection taking into account associated population/species-specific covariates, using BayPass v2.1 (http://www1.montpellier.inra.fr/CBGP/software/baypass/), for the putative duplications in the Heliconius Set (Gautier, 2015). The duplication events were considered as dominant binary markers. We used country coordinates and species as population-specific covariates. The covariates were defined as follows: Costa Rica: 9.7489, 83.7534; Panama: 8.5380, 80.7821; French Guiana: 3.9339, 53.1258; *H.* cydno: 1 and H. *melpomene*: 2. Under the Standard Covariate Model we estimated for each duplication event the Bayes Factor, the empirical Bayesian *P*-value and its underlying regression coefficient using an Importance Sampling algorithm. We simulated the data under the Inference Model to calibrate the neutral distribution of XtX. XtX was used to identify loci subjected to adaptive divergence. After calibrating XtX we ran the Markov chain Monte Carlo algorithm using posterior estimates available from

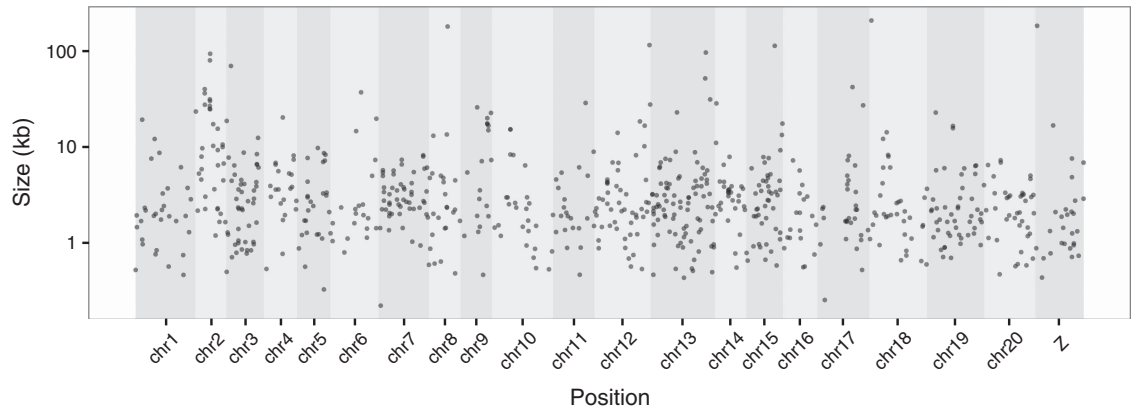


Figure 4 Distribution of the *Heliconius* duplication set mapped to the Hmel2 reference genome. *H. cydno* and *H. melpomene* genotyping sets were filtered and exclude duplications with a median read count of >500 reads per sample or not genotyped in any of the samples. The two high-quality genotyping sets were merged to produce the Heliconius duplication set (Heliconius Set, Figure 1a and Table 1). Each putative duplication on the Heliconius set is represented by a point according to position in the genome (x axis) and size (kb).

the previous analysis and we corrected for location using just one covariable at a time, as suggested by Gautier (2015). Finally, we selected the duplication events that had observed XtX estimates above the 98% threshold of the simulated data (XtX >7.9). We cross-referenced the regions selected from BayeScan and BayPass analyses to look for overlaps between the two methods.

RESULTS

Duplication maps for *H. cydno* and *H. melpomene*

We identified a Discovery duplication set of 1920 putative *H. cydno* duplications and 1591 putative *H. melpomene* duplications (Table 1, Discovery set: merged by species) based on whole-genome resequencing data from 10 wild *H. cydno* samples and 10 wild *H. melpomene* samples (Kronforst *et al.*, 2013; Supplementary Table S1). We genotyped the discovery sets in a further 10 *H. melpomene* and 4 *H. cydno* samples (Martin *et al.*, 2013). After removing duplications with low-quality genotypes and high RD and duplications where all samples differed from the *H. melpomene* reference genome, we retained 497 putative *H. cydno* duplications and 463 *H. melpomene* duplications (Table 1, Genotyping set; Figure 4 and Supplementary Figures S5 and S6). We then merged redundant duplications in the *H. cydno* and *H. melpomene* Genotyping Sets, where two variants overlapped in over 90% of their total length, to produce the Heliconius Set containing 744 duplications ranging in size from 228 bp to 207 510 bp (median 5693 bp) (Table 1, Heliconius set; Supplementary Figures S7–S9).

Validation rate as estimated by analysis of PacBio single-molecule long reads

We validated our pipeline using Illumina and PacBio sequencing data for a single chromosome from two families of *H. melpomene* and *H. cydno*. We first ran our pipeline on the Illumina data for chromosome 2 and then validated the calls using the PacBio data. Using the Illumina sequenced trio, we identified 97 duplications on chromosome 2 in *H. melpomene* and 137 in *H. cydno* after filtering. We validated 96.9% of the *H. melpomene* and 95.6% of the *H. cydno* calls using single-molecule PacBio SRs for each species separately. We also ran the Heliconius Set of duplications using the same PacBio data, combining the data from *H. cydno* and *H. melpomene*. This confirmed 65.5% of putative duplications. The lower validation rate on the Heliconius Set duplications is because of the fact that these are different individuals and populations compared with our PacBio data. In the Heliconius set a third to a quarter of all duplications identified only occurred in a single individual and hence were unlikely to be present in the PacBio data (Supplementary Figure S8). Nonetheless, the high validation observed in our reference trios suggests that our pipeline is correctly identifying duplications from Illumina data.

Effect of genome structure on duplication distribution

Most duplications occurred in a small number of samples and there were only a few duplications at high frequency among all the samples (Supplementary Figure S8). For example, in the *H. cydno* genotyping set, 26.8% of the duplications are singletons and, in the *H. melpomene* 32.5%. The number of duplications per chromosome in the Heliconius Set is not equally distributed along the different chromosomes (Supplementary Figure S9A) and is weakly correlated with chromosome size ($r^2 = 0.344$; Supplementary Figure S9B). There was also variation between individual chromosomes in the number of duplications per Mb ($F(20,723) = 14.2$, $P < 0.001$). Chromosome 18 tended to have fewer duplications, whereas chromosome 17 showed an excess of duplications per Mb compared with other chromosomes (*post hoc* Tukey's HSD (honest significant difference) test with correction for multiple testing). We did not observe any excess or depletion of duplication events towards the centres of chromosomes in the Heliconius Set (Supplementary Figure S10).

Principal component analysis of the genotyped *H. cydno* and *H. melpomene* sets

We tested for population structure in the Heliconius Set of duplications genotyped as co-dominant markers using principal component analysis. In total, 17.57% of the total variance was explained by the first two principal components (PCs; PC1 12.97% and PC2 4.6%). Along PC1 the samples separated by species and geography (Figure 3), with all populations distinct except *H. m. melpomene* and *H. m. rosina* samples from Panama that are known to be genetically very similar (Martin *et al.*, 2013). However, PC2 separates the Costa Rica samples from those from Panama and French Guiana. It seems most likely that this is a methodological artefact because samples from different countries came from different sequencing runs (Supplementary Table S1). In addition, our call set was generated from the Costa Rica data set, and subsequently genotyped on both sample sets. Within Costa Rica, PCA analyses separate populations by geography and species as expected (Supplementary Figure S11).

Table 2 Functional impact of the Heliconius set

Species	*Complete gene*	*%*	*< Sim 2.5%*	*Gene*	*%*	*<Sim 2.5%*	*CDS*	*%*	*<Sim 2.5%*	*Intron*	*%*	*<Sim 2.5%*	*UTR*	*%*	*<Sim 2.5%*
Heliconius melpomene	23	5.2	No	157	35.3	Yes	92	20.7	Yes	45	10.1	No	27	6.1	No
Heliconius cydno	41	8.9	No	210	45.8	No	154	33.6	No	42	9.2	No	20	4.4	No

Abbreviations: CDS, coding sequence; UTR, untranslated region.
Observed absolute counts and proportion of duplications overlapping complete genes, genes, CDS, introns and UTRs. <Sim 2.5% column indicates whether the observed proportion of overlap with each category falls within the 2.5% confidence interval of the simulated data overlap after 10 000 iterations. If <sim 2.5% is 'No', then duplication counts are not within the 2.5% confidence interval and the overlaps observed do not significantly differ from random expectations. If 'Yes', then counts are within the 2.5% confidence interval and the overlap observed is significantly less than expected under a random distribution. A single duplication can fall into several subcategories.

Overlap between duplication and genes

We found that the genotyped duplications in *H. melpomene* overlapped with genes and CDSs significantly less often than expected by chance, whereas the rate of overlap with UTRs and introns did not differ from the null expectation under a random distribution (Table 2 and Supplementary Figure S12). This is consistent with the idea that duplications involving functional regions have a greater probability of being deleterious, and are therefore more likely to be removed by selection. In contrast to *H. melpomene*, in *H. cydno*, there was no significant deviation from the null expectation in the rate of overlap between genotyped duplications and genes, CDSs, UTRs or introns.

Enrichment of biological functions in the Heliconius Set

The duplications we have identified are not equally distributed across the genome (Figure 4 and Supplementary Figure S9). The heterogeneity observed across the landscape is likely to be a reflection of biases in the rates at which duplications arise in certain regions or a bias in the preservation of duplications in specific functional classes because of the action of natural selection. It has been shown that multigene families, specifically those involved in environmental responses, are particularly prone to being duplicated/retained (Duvaux *et al.*, 2015). We detected 19 gustatory receptors that had been previously identified as putatively duplicated by CNVnator analysis (Briscoe *et al.*, 2013). Moreover, we tested whether any biological functions were overrepresented in the Heliconius set of duplications using PANTHER (Supplementary Figure S13). Within the *Heliconius* set there were 1710 different family classes of which 1181 were associated with predicted biological processes. Of these processes, 26 different biological function categories were identified as overrepresented in the *Heliconius* set based on the *D. melanogaster* reference list ($P<0.005$) (Supplementary Figure S13 and Supplementary Table S2). These were involved in transketolase, phosphatase, endodeoxyribonuclease, metallopeptidase, lipid transport, deacetylase, oxidoreductase and transferase activity. There was also a set of 529 family classes that are overrepresented in the Heliconius set but do not have a specific Gene Ontology (GO) term, biological or specific molecular function associated with them but include ejaculatory bulb-specific protein, male sterility protein, cuticle formation and transposable element related (Supplementary Figure S13, Unclassified; Supplementary Table S3). Structural constituents of the cytoskeleton, protein binding, DNA binding transcription factor and kinase activity were molecular function categories underrepresented in the *Heliconius* set. The biological function that was most overrepresented in the entire set was the GO category related to the pentose-phosphate shunt (primary metabolic process, fold enrichment 18.35, $P=5.4e{-}07$). Immune system processes were underrepresented in our set (fold enrichment <0.2, $P=2.59e{-}04$).

Identification of outlier duplications in the Heliconius Set potentially under selection

To characterize patterns of divergence observed between *H. melpomene* and *H. cydno* we first calculated F_{ST} between the two species and identified candidate outlier regions using BayeScan for the Heliconius Set of duplications, treating putative duplications as co-dominant (presence/absence) markers. After correcting for false positives we found nine duplications that are candidates for selection (Supplementary Figure S14A and Supplementary Table S4). We also ran BayPass that conducts a similar test by accounting for sample location and species. This produced six putative duplicated regions above the simulated significance threshold (Supplementary Figure S14B and Supplementary Table S4), four of which were also identified by BayeScan (Table 3). We consider the four outlier events found by both tests to be strong candidates for directional selection. One region, on chromosome 15, is located in an intergenic region upstream of the gene *cortex* that is involved in the regulation of yellow and white wing pattern elements (Figure 1b) (Nadeau *et al.*, 2016). The other three regions overlap with genes, predicted to be a Kazal-type serine protease (chromosome 9), an odorant binding protein (chromosome 18) and a regulator of the cell cycle and nitrogen compound metabolic processes (chromosome 21) (Table 3). All four candidate selected duplications are absent in the *H. melpomene* samples and present in 13 or 14 of the 14 *H. cydno* samples.

DISCUSSION

Gene duplication is an important source of genetic fuel for evolutionary diversification, and can also contribute to speciation. Here we have used short-read genome sequence data to identify signatures of CNV in natural populations. We have used single-molecule sequencing to validate our pipeline, with a validation rate of ~96% within families. We have successfully identified 744 loci and genotyped them (presence/absence) in 34 wild individuals sampled from the two species *H. melpomene* and *H. cydno*.

Despite the ubiquitous nature of duplications, different chromosomes might be expected to contribute differently to the overall duplication landscape. Large chromosomes tend to have the highest absolute duplication counts but chromosome size is not the sole predictor of duplication distributions. Sex chromosomes, which have more repetitive content, smaller population sizes and lower levels of background selection than autosomes, have been shown to have a higher duplication load per base pair than autosomes in *D. simulans* and in *D. melanogaster* (Charlesworth, 2012; Mackay *et al.*, 2012; Zichner *et al.*, 2013; Rogers *et al.*, 2014, 2015). However, the X chromosome of *Drosophila yakuba* does not contain an excess of duplications compared with the autosomes and no signals of adaptation through duplication have been identified. Similarly, the *Heliconius* duplication set does not harbour an excess of duplications on the Z chromosome compared with the autosomes. It is possible that duplications are more difficult to detect on the Z chromosome that

Table 3 Putative duplicated loci under selection between *Heliconius cydno* and *Heliconius melpomene*

Chr	*Scaffold*	*Start*	*End*	*Size*	*BayeScan log10(PO)*	*BayPass mean XtX*	*Freq in H. melpomene*	*Freq in H. cydno*	*PANTHER GO-Slim Biological process*	*Hmel2 annotation*
9	Hmel209007	4 344 840	4 364 959	20 119	1.7222	7.95239143	0	0.93	Kazal-type serine protease inhibitor	HMEL009267
15	Hmel215006	1 190 144	1 196 212	6068	1.8414	8.78515118	0	1	NA	upstream of *cortex*
18	Hmel218003	221 730	42 9239	207 509	1.894	8.75630075	0	1	Protein targeting	OBP41
									Intracellular protein transport	HMEL013558
									Transport	HMEL013559
									Localization	HMEL003174
									Biological regulation	HMEL003175
									Asymmetric protein localization	HMEL003862
										HMEL003863
21	Hmel221012	779 541	796 444	16 903	1.72	8.35788884	0	0.93	Regulation of the cell cycle	HMEL016617
									Regulation of biological process	HMEL016621
									Porphyrin-containing compound Metabolic process	HMEL016620
									Nitrogen compound metabolic process	
									Regulation of translation	
									Primary metabolic process	
									mRNA transcription	
									Nucleobase-containing compound metabolic process	
									Cell differentiation, developmental process	
									Regulation of transcription from RNA pol II promoter	

Abbreviation: NA, not available.
The four duplications in the Heliconius set identified as outliers by BayeScan and BayPass analysis. Chromosome position, scaffold name, start, end and size of each putative duplication are indicated. log10 (Posterior Probabilities) from the BayeScan analysis is indicated per duplication between the *H. melpomene* and *H.* cydno. All these loci had positive values of α that suggests diversifying selection. BayPass XtX mean for each loci is also indicated for each species after correcting for location. Allele frequencies calculated as co-dominant markers are shown for each species at the loci (genotyped by Delly2). PANTHER GO-Slim biological processes and Hmel2 annotations retrieved from Hmel2.gff (Davey *et al.*, 2016).

has higher divergence than the rest of the genome (Martin *et al.*, 2013) and higher proportion of repetitive content (Conrad and Hurles, 2007). Further work will be needed to compare the landscape of duplications across sex chromosomes.

Duplications are not homogenously distributed across the genome (Figure 2 and Supplementary Figures S5 and S6). There was no bias towards telomeric regions as has been documented for humans (Zhang *et al.*, 2005). *Heliconius*, like *C. elegans*, have holocentric chromosomes and, to our knowledge the enrichment of structural variations in telomeric regions (and/or pericentrimeric regions) has yet to be documented for organisms with this chromosomal organization (Farslow *et al.*, 2015). The number of singletons identified in our data set (a quarter to a third of all duplications) is on the same order of magnitude as that seen previously. For example, Duvaux *et al.* (2015) reported 31% singletons in pea-aphid clones.

A large proportion of structural variants arising in genomes are slightly or moderately deleterious and therefore experience purifying selection (Emerson *et al.*, 2008; Zichner *et al.*, 2013). In *D. melanogaster*, fewer duplications were found in coding sequence as compared with random expectation (Zichner *et al.*, 2013). Consistent with this, we found that in the *H. melpomene* Genotyping Set duplications are biased away from coding regions, although they are not biased away from or towards intronic or UTR regions. However, we did not find a similar bias in *H. cydno*, and saw no significant depletion of the number of duplications in *H. cydno* as compared with *H. melpomene*. This goes against expectations, given that the effective population size of H. *cydno* has been inferred to be around four times greater than that of *H. melpomene* (Kronforst *et al.*, 2013), consistent with the significantly higher genome-wide heterozygosity in *H. cydno* (Martin *et al.*, 2013). Therefore, we might expect selection to operate more effectively and duplications to be more efficiently removed from *H. cydno*, but this does not appear to be the case. We do not have any good explanation for this.

Although most structural variants may be deleterious, there is particular interest in those few that have positive effects. There are now many examples in which gene duplicates provide the genetic fuel for adaptation, and have been shown to be under positive selection (Beisswanger and Stephan, 2008; Arroyo *et al.*, 2012; Blount *et al.*, 2012). Here, we are specifically interested in speciation. Gene duplicates have been implicated in reproductive isolation for both animals and plants. For example, the Odysseus gene that causes hybrid sterility between *D. mauritiana* and *D. simulans* is a duplicate of the *unc-4* gene (Ting *et al.*, 2004). In *A. thaliana*, paralogues of an essential duplicate gene that evolved divergently interact epistatically in some interspecific crosses and control a recessive embryo lethality (Bikard *et al.*, 2009). In the context of *Heliconius*, we are specifically interested in speciation and divergent selection between the closely related species, *H. melpomene* and *H. cydno*. Using BayeScan and BayPass we identified a relatively small number of duplications that are putatively divergently selected between these species.

Many functionally important regions in different genomes have been documented to evolve through gene duplication followed by neo or subfunctionalization. Genes responsible for environmental response are known to be overrepresented as duplicated sequences in a range of organisms from humans to fruit flies and butterflies (Johnson *et al.*, 2001; Tuzun *et al.*, 2005; Hahn *et al.*, 2007; Briscoe *et al.*, 2013) and in line with previous studies we have detected an enrichment of genes involved in sensory perception (Briscoe *et al.*, 2013; Rogers *et al.*, 2014;

Duvaux *et al.*, 2015; Paudel *et al.*, 2015). For example, we detected gustatory receptors that had already been identified in *Heliconius* (Briscoe *et al.*, 2013) but we also detected others such as olfactory receptors and olfactomedin-related proteins (Supplementary Table S3). Specifically, in our outlier analysis there is an odorant binding protein that is divergent in copy number between *H. cydno* and *H. melpomene* (OBP41, Table 3). Several hypotheses have been put forward to explain the trend of increased CNV among genes involved in environmental response. On one hand, these CNVs might be maintained by positive selection as outlier analysis-based methods have shown an enrichment for these GO classes (Duvaux *et al.*, 2015; Paudel *et al.*, 2015; Rogers *et al.*, 2015). On the other hand, these differences could occur simply because certain sequence motifs like non-B DNA forming sequence are more common in gene-rich regions and, at the same time, they increase the rate of CNV formation (Sjödin and Jakobsson, 2012). Gene categories overrepresented in CNV are also enriched within segmental duplications, and segmental duplications are very structurally dynamic (Conrad and Hurles, 2007). Moreover, families with multiple paralogues are more prone to further copy number variation (Hastings *et al.*, 2009).

Not all the putative duplications we found as outliers were involved in environmental response. Another candidate locus under divergent selection was found near the *cortex* gene that controls the yellow hindwing bar and white/yellow forewing patterns that differ between *H. m. rosina* and *H. cydno* (Nadeau *et al.*, 2016). Moreover, we have also found an enrichment of male reproductive proteins in the Heliconius Set (Supplementary Table S3). These proteins evolve rapidly and are commonly duplicated in, for example, *D. yakuba* (Rogers *et al.*, 2014). It was somewhat surprising, however, that we did not observe an enrichment for immunity-related genes.

Interestingly, the four putative duplicated regions we have identified as excessively differentiated in *H. cydno* and *H. melpomene* were all nearly fixed in *H. cydno* but not in *H. melpomene*. *H. melpomene* and *H. cydno* differ in many aspects of their ecology and behaviour. Shifts in host plant have played a central role in their diversification. The evolution of host-use strategies reflects a tradeoff between selection pressures (Merrill *et al.*, 2013). For example, gene duplications that persist in an evolving lineage have often been found to be beneficial because of a protein dosage effect in response to environmental conditions. Host-plant systems may be subject to rapid coevolution and duplicated loci in *H. cydno* could be related to the fact that *H. cydno* is a host plant generalist and *H. melpomene* is a specialist (Merrill *et al.*, 2013).

The duplications we have identified as being under selection between *H. cydno* and *H. melpomene* may play a role in species divergence. We have shown that, despite being ubiquitous, the landscape of duplications in *Heliconius* is heterogeneous and likely to be under both positive and negative selection. The putative duplications we found merit further investigation for their potential role in host plant and mate recognition differences between the species.

DATA ARCHIVING

All short-read sequence data are publicly available (Kronforst *et al.*, 2013; Martin *et al.*, 2013; Malinsky *et al.*, 2016). Long-read Pacific Biosciences data are available at European Nucleotide Archive accession PRJEB6424. Custom scripts, Genotyping Sets and Heliconius Set are available from Dryad (doi:10.5061/dryad.8jv30).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

AP is funded by a NERC studentship (PFZE/063). CDJ, SLB, JWD and SHM are funded by ERC grant SpeciationGenetics (Grant Number 339873). Pacific Biosciences sequencing was carried out by Karen Oliver in collaboration with Richard Durbin at the Sanger Institute, supported by European Research Council (ERC) Grant Number 339873, Wellcome Trust Grant Number 098051. We thank Jenny Barna and Stuart Rankin for computing support. Analyses were carried out using the Darwin Supercomputer of the University of Cambridge High Performance Computing Service (http://www.hpc.cam.ac.uk/), provided by Dell Inc. using Strategic Research Infrastructure Funding from the Higher Education Funding Council for England, and funding from the Science and Technology Facilities Council. We thank the editor and three anonymous reviewers for their comments that helped us to improve this manuscript.

Abyzov A, Urban AE, Snyder M, Gerstein M (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**: 974–984.

Armengol L, Villatoro S, Gonzalez JR, Pantano L, Garcia-Aragones M, Rabionet R *et al.* (2009). Identification of copy number variants defining genomic differences among major human groups. *PLoS One* 4: e7230.

Arroyo JI, Hoffmann FG, Opazo JC (2012). Gene duplication and positive selection explains unusual physiological roles of the relaxin gene in the European rabbit. *J Mol Evol* **74**: 52–60.

Beisswanger S, Stephan W (2008). Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in Drosophila. *Proc Natl Acad Sci USA* **105**: 5447–5452.

Bikard D, Patel D, Le Mette C, Giorgi V, Camilleri C, Bennett MJ *et al.* (2009). Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. *Science* **323**: 623–626.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE (2012). Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* **489**: 513–518.

Briscoe AD, Macias-Munoz A, Kozak KM, Walters JR, Yuan F, Jamie GA *et al.* (2013) Female behaviour drives expression and evolution of gustatory receptors in butterflies. *PLoS Genet* **9**: e1003620.

Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, Kalbe M *et al.* (2014). Extensive copy-number variation of young genes across stickleback populations. *PLoS Genet* **10**: e1004830.

Charlesworth B (2012). The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the Drosophila X chromosome. *Genetics* **191**: 233–246.

Cheang CC, Tsang LM, Chu KH, Cheng I-J, Chan BKK (2013). Host-specific phenotypic plasticity of the turtle barnacle Chelonibia testudinaria: a widespread generalist rather than a specialist. *PLoS One* **8**: e57592.

Chen K, Chen L, Fan X, Wallis J, Ding L, Weinstock G (2014). TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res* **24**: 310–317.

Conant GC, Wolfe KH (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* **9**: 938–950.

Conrad DF, Hurles ME (2007). The population genetics of structural variation. *Nat Genet* **39**: S30–S36.

Davey JW, Chouteau M, Barker SL, Maroja L, Baxter SW, Simpson F *et al.* (2016). Major improvements to the Heliconius melpomene genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. *G3 (GenesGenomes Genet)* **6**: 695–708.

Deng C, Cheng C-HC, Ye H, He X, Chen L (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc Natl Acad Sci USA* **107**: 21593–21598.

Duvaux L, Geissmann Q, Gharbi K, Zhou J-J, Ferrari J, Smadja CM *et al.* (2015). Dynamics of copy number variation in host races of the pea aphid. *Molec Biol Evol* **32**: 63–80.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* **320**: 1629–1631.

Farslow JC, Lipinski KJ, Packard LB, Edgley ML, Taylor J, Flibotte S *et al.* (2015). Rapid Increase in frequency of gene copy-number variants during experimental evolution in *Caenorhabditis elegans*. *BMC Genom* **16**: 1044.

Feuk L, Carson AR, Scherer SW (2006). Structural variation in the human genome. *Nat Rev Genet* **7**: 85–97.

Feulner PGD, Chain FJJ, Panchal M, Eizaguirre C, Kalbe M, Lenz TL *et al.* (2013). Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. *Molec Ecol* **22**: 635–649.

Foll M, Gaggiotti O (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**: 977–993.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Gautier M (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* **201**: 1555–1579.
Hahn MW, Han MV, Han S-G (2007). Gene family evolution across 12 Drosophila genomes. *PLoS Genet* **3**: e197.
Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009). Mechanisms of change in gene copy number. *Nat Rev Genet* **10**: 551–564.
Hunt DM, Dulai KS, Cowing JA, Julliot C, Mollon JD, Bowmaker JK *et al.* (1998). Molecular evolution of trichromacy in primates. *Vision Res* **38**: 3299–3306.
Iskow RC, Gokcumen O, Lee C (2012). Exploring the role of copy number variants in human adaptation. *Trends Genet* **28**: 245–257.
Jiggins CD, Naisbit RE, Coe RL, Mallet J (2001). Reproductive isolation caused by colour pattern mimicry. *Nature* **411**: 302–305.
Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M *et al.* (2001). Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**: 514–519.
Jombart T, Ahmed I (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**: 3070–3071.
Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA (2009). Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. *Genome Res* **19**: 1404–1418.
Katju V (2012). In with the old, in with the new: the promiscuity of the duplication process engenders diverse pathways for novel gene creation. *Int J Evol Biol* **2012**: 341932–24.
Katju V, Bergthorsson U (2013). Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front Genet* **4**: 273.
Keane OM, Toft C, Carretero-Paulet L, Jones GW, Fares MA (2014). Preservation of genetic and regulatory robustness in ancient gene duplicates of *Saccharomyces cerevisiae*. *Genome Res* **24**: 1830–1841.
Kronforst MR, Hansen MEB, Crawford NG, Gallant JR, Zhang W, Kulathinal RJ *et al.* (2013). Hybridization reveals the evolving genomic architecture of speciation. *Cell Rep* **5**: 666–677.
Layer RM, Chiang C, Quinlan AR, Hall IM (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
Li H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN].
Lin K, Smit S, Bonnema G, Sanchez-Perez G, de Ridder D (2015). Making the difference: integrating structural variation detection tools. *Brief Bioinform* **16**: 852–864.
Lunter G, Goodson M (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* **21**: 936–939.
Lynch M, Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.
Lynch M, Force A (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**: 459–473.
Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D *et al.* (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**: 173–178.
Malinsky M, Simpson JT, Durbin R (2016). trio-sga: facilitating de novo assembly of highly heterozygous genomes with parent-child trios. *bioRxiv*.
Manzanares M, Wada H, Itasaki N, Trainor PA, Krumlauf R, Holland PW (2000). Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head. *Nature* **408**: 854–857.
Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F *et al.* (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res* **23**: 1817–1828.
McMillan WO, Jiggins CD, Mallet J (1997). What initiates speciation in passion-vine butterflies? *Proc Natl Acad Sci USA* **94**: 8628–8633.
Merrill RM, Dasmahapatra KK, Davey JW, Dell'Aglio DD, Hanly JJ, Huber B *et al.* (2015). The diversification of Heliconius butterflies: what have we learned in 150 years? *J Evol Biol* **28**: 1417–1438.
Merrill RM, Naisbit RE, Mallet J, Jiggins CD (2013). Ecological and genetic factors influencing the transition between host-use strategies in sympatric Heliconius butterflies. *J Evol Biol* **26**: 1959–1967.
Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD (2016). PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* **44**: D336–D342.
Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C *et al.* (2011). Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59–65.
Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R *et al.* (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* **43**: D213–D221.
Muñoz AG, Salazar C, Castano J, Jiggins CD, Linares M (2010). Multiple sources of reproductive isolation in a bimodal butterfly hybrid zone. *J Evol Biol* **23**: 1312–1320.
Nadeau NJ, Pardo-Diaz C, Whibley A, Supple MA, Saenko SV, Wallbank RWR *et al.* (2016). The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature* **534**: 106–110.
Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW *et al.* (2011). Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philos Trans R Soc B* **367**: 343–353.
Ohno S (1970). *Evolution by Gene Duplication*. Springer-Verlag: New York, NY, USA.
Paudel Y, Madsen O, Megens H-J, Frantz LAF, Bosse M, Crooijmans RPMA *et al.* (2015). Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. *BMC Genom* **16**: 330.
Qian W, Zhang J (2014). Genomic evidence for adaptation by gene duplication. *Genome Res* **24**: 1356–1362.
Quinlan AR, Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333–i339.
Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR (2014). Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. *Molec Biol Evol* **31**: 1750–1766.
Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR (2015). Tandem duplications and the limits of natural selection in *Drosophila yakuba* and *Drosophila simulans*. *PLoS One* **10**: e0132184.
Sjodin P, Jakobsson M (2012). Population genetic nature of copy number variation. *Methods Mol Biol* **838**: 209–223.
Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**: 2032–2034.
Tattini L, D'Aurizio R, Magi A (2015). Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng Biotechnol* **3**: 92.
Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* **28**: 2711–2718.
The Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**: 94–98.
Ting C-T, Tsaur S-C, Sun S, Browne WE, Chen Y-C, Patel NH *et al.* (2004). Gene duplication and speciation in Drosophila: evidence from the Odysseus locus. *Proc Natl Acad Sci USA* **101**: 12232–12235.
Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM *et al.* (2005). Fine-scale structural variation of the human genome. *Nat Genet* **37**: 727–732.
Yao W (2015). intansv: Integrative analysis of structural variations. *R package version 1.9.2*.
Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865–2871.
Zhang L, Lu HHS, Chung W-Y, Yang J, Li W-H (2005). Patterns of segmental duplication in the human genome. *Molec Biol Evol* **22**: 135–141.
Zhao M, Wang Q, Wang Q, Jia P, Zhao Z (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinform* **14**(Suppl 11): S1.
Zichner T, Garfield DA, Rausch T, Stutz AM, Cannavo E, Braun M *et al.* (2013). Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. *Genome Res* **23**: 568–579.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)

Appendix B

Protocol for dissections of the reproductive tract for total RNA extraction

Protocol for dissections of the reproductive tract for RNA extraction

A. Materials

1. RNAlater
2. RnaseZAP
3. 70% Ethanol
4. A pair of fine dissecting scissors
5. Two small dissection plates (preferably with silicon base)
6. Fine point dissection scissors
7. Flat soft tweezers (to hold wings)
8. Scissors for cutting the wings & separate abdomen from thorax
9. Pins
10. Micro-ruler
11. Camera adapted to take pictures in a microscope
12. 0.5ml screw-cap tubes with RNAlater (THORAX, ABDOMEN, GUT, BURSA, OVARIES)

B. Methods

1. Clean work surface and tools very well

2. Place dissecting plate under the microscope and fill 1/3 full with RNAlater

a. Do not fill the plate to the ream. The solution has to be just enough to cover the tissue.

3. Hold the butterfly with the flat tweezers and cut the abdomen off and let it *fall* into the dissection plate with the RNAlater
    a. After this step the abdomen should be in the plate with the RNAlater. We should be left with a butterfly with the thorax and the wings

4. Cut the wings off and put the THORAX in the first tube with RNAlater
    a. The abdomen is not going to sink in the RNAlater from the plate without being pinned. Cut the wings and preserve the thorax, pin the abdomen after the butterfly is dead.

5. Pin the abdomen to the silicone so that the whole abdomen is submerged when cut open
    a. The dorsal side of the abdomen should be facing upwards and the pins should be placed as closer to the edge as possible.
    b. Stretch as much as possible.

6. Cut a straight line with the fine point scissors as close to the scales as possible
    a. Cut as close to the scales as possible. Start by inserting the edges of the scissors within one of the openings then cut through the whole abdomen (Figure 1).

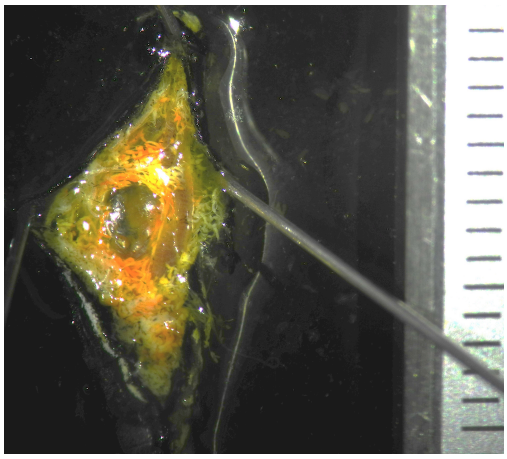


Figure 1. Overview of the cut abdomen of a female before dissecting inside

Scale in mm

7. **With the pair of fine tweezers start *peeling* the mass of fat bodies (yellow) from the abdomen**
    a. The reproductive tract will be opposite the slit that we cut under the “air sac”. It will be surrounded by tracheoles.
    b. After this step there will be a “shell” – mainly just the black abdomen – and a large mass surrounded by a *cloud* of fat bodies.

8. **Remove the GUT and store it**
    a. The gut is large and it comes attached to the end of the vulva.
    b. Gently pull it with a pair of tweezes from the rest.
    c. Remove the fat bodies attached to it to store it.

d. Keep all the fat bodies to one side of the plate – do not discard them – to store in the ABDOMEN tube.

9. Locate the bursa and clean it

a. At this point it should be easy to see where the bursa is. Walk from the bursa and find the connection to the ovaries.

b. Start by cleaning the bursa gently with the tweezers.

10. Clean the ovaries and the ovarioles

11. Place the OVARIES, the BURSA and the GUT in a clean dissection plate

a. After cleaning a new dissection plate fill it by 1/3 with RNAlater. Pick the ovaries & bursa from the first dissection plate and place them in the clean one.

12. Take pictures of the reproductive tract

a. Include a scale (Figure 2).

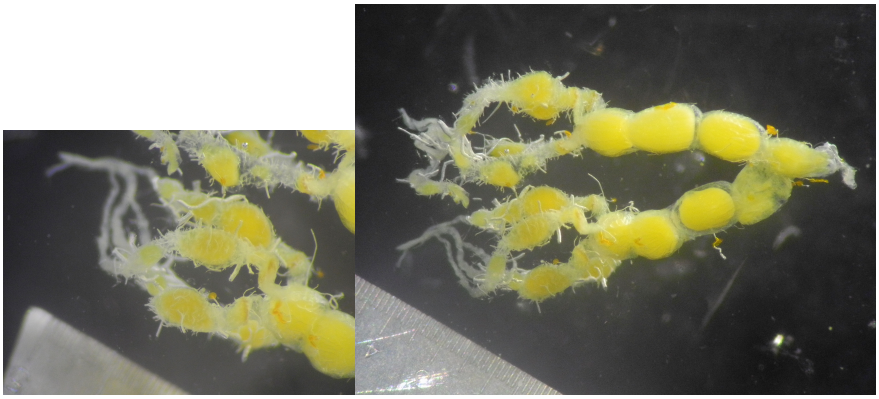


Figure 2. Example pictures of dissected fertile

13. Store the OVARIES and the BURSA separately in RNAlater

14. Transfer the GUT to the clean dissection plate and take pictures

15. Store the GUT and the ABDOMEN (the fat bodies and the black *shell*) in separate RNA later tubes

16. Move the tubes to the fridge for 24h

17. After 24h in the fridge move the tubes with the tissue to the freezer
    a. Freeze and thaw as LITTLE as possible to prevent degradation.

Appendix C

Total RNA extraction protocol for mRNA sequencing

Total RNA extraction protocol for mRNA sequencing

RNA extraction: Guanidium thiocyanate-phenol-chloroform combined with silica matrix protocol

Equipment and reagents

1. TRIzol
2. Chloroform (>99% containing amylenes)
3. 70% ethanol in DEPC water – COLD
4. RNAse-free water (DEPC treated)
5. RNeasy plus mini kit
6. DNAseI (Ambion)
7. Sterile forceps
8. Sterile, RNase-free pipet tips
9. Disposable gloves
10. Refrigerate centrifuge (with rotor for 2ml tubes)
11. Micro-centrifuge (with rotor for 2ml tubes)
12. Vortex
13. Fume hood
14. Heat block

Before starting

Add ethanol to Buffer RPE

SET UP

For work on the fume hood

a. Set up the centrifuge for 14000rpm at 4C
b. Aliquot of TRIzol – 1ml per sample
c. 2mL eppendorf – two per sample
d. Timer
e. 70% ethanol in falcon tube – 600ul per sample – COLD
f. Aliquot chloroform – 200ul per sample
g. Pipettes: P200, P1000, P20
h. Tweezers
i. RNaseZAP
j. Tissue homogenization RNA balls (stainless steal)
k. RNeasy spin columns – 1 per sample
l. 2ml collection tubes – 3 per sample
m. Set up centrifuge for 14000rpm
n. Wipes
o. Ice bucket

For work in bench

a. Set up centrifuge for 10000rpm
b. Buffer RW1 – 700ul per sample
c. Buffer RPE – 500ul per sample
d. Buffer RPE – 500ul per sample
e. 1.5ml eppendorf – two per sample
f. RNase-free water – 30ul-50ul per sample
g. Pipettes: P100, P1000
h. RNaseZAP
i. 0.5ml RNase free tubes – 3 per sample (for QC)
j. Set up heat block to 37ºC

k. Ice bucket

- - - - - - - - - - - - - EXTRACTION IN THE HOOD - - - - - - - - - - - -

Sample preparation

1. The tissue is immersed in RNAlater

   Use a wipe to remove excess and immediately add tissue to a 2 ml tube containing **1ml of TRIzol** using sterile forceps

   Note: 1ml of TRIzol is appropriate for 50-100mg of tissue

Homegenization

1. Homogenise the starting material using either the Polytron PT1600E (in S9) or the Tissue lyser (F106)

   ### a. Polytron

   1. Wash the drill for 20secs for each wash:
      2 times with SDS,
      1 time with EtOH 100%,
      1 time with RNAzap,
      2 times with DEPC water

   2. Homogenise the sample for 1 minute slowly increasing to ~20k rpm

      Verify absence of bubbles, and if it happens, repeat washes and

homogenisation

b. Tissue lyser

1. Place one bead per tube with the right amount of tissue and Trizol
2. Homogenise at frequency 30 for 2 minutes (program 2)
3. Check the bottom of the tubes and if it has cracked during the homogenisation, transfer the solution to a new tube with a bead
4. Repeat the homogenisation at frequency 30 for 2 minutes

Note: First run program 4 (10 s) on the tissue lyser to make sure the machine is well set up. The two screws turn opposite ways, and both Quiagen symbols need to face the same way. For better homogenization of the material change orientation of the Quiagen symbols after the first 2min homogenization. There is only one way the eppendorf tubes will fit the blocks – the lids have to be facing the plastic ridges on the blocks.

2. Incubate for **5 min RT**

- - - - - - - - - - - - - KEEP TUBES IN ICE FROM NOW - - - - - - - -

Phase separation

1. Add **200 uL of chloroform** and **vortex for 15 s** (0.2mL chloroform per mL of TRIzol)
2. Incubate the tube for **3 min at RT**
3. Spin at 14 000 rpm for 15 min at 4ºC
4. **Carefully** remove aqueous phase (top) and transfer to a new tube (~580 uL)

It is **extremely important** not to get any of the material from the aqueous/organic interface. It is suggested to sacrifice aqueous material rather than risk taking the precipitate

RNA precipitation

1. Measure the volume of the aqueous phase
2. **SLOWLY** add an equal volume of 70% EtOH

   Slow introduction of EtOH is important to avoid localised precipitation of RNA

3. Mix by pipetting as it is added

RNA purification on matrix (using RNeasy plus mini kit, Qiagen)

From step 4 of the kit handbook

Each 2mL eppendorf had ~ 1 mL TRIzol + 200 ul CHCl3 + 600ul 70% EtOH = **1800ul**

1. Transfer up to **700 ul of the sample**, including any precipitate that might have formed to an RNeasy spin column placed in a 2mL collection tube
2. Centrifuge for **30 s at > 10 000 rpm** (8,000 g)
3. Discard the flow-through
4. Reuse the collumn if the sample volume exceeds 700 uL (i.e. repeat ~2 times)
5. Centrifuge successive aliquots in the same RNeasy spin column
6. Discard flow-through after each centrifugation

Note: If the expected yield is larger than the RNA-binding capacity of the column (100 ug for RNeasy) the sample should be split and purified using multiple columns

- - - - - EXTRACTION CONTINUED IN THE LAB (OUT OF THE HOOD) - - - -

A. Take DNAse I Ambion's buffers and let them defrost on ice

RNA purification on matrix (using RNeasy plus mini kit, Qiagen)

7. Add **700 ul Buffer RW1** to the RNeasy spin column
8. Centrifuge for **30 s at >= 10 000 rpm** to wash the spin column membrane
9. Discard the flow-through
10. Transfer column into a new 2mL collection tube
11. After centrifugation carefully remove the RNeasy spin column from the collection tube so that the column does not contact the flow-through
12. Add **500 ul Buffer RPE** to the RNeasy spin column
13. Centrifuge for **30 sec at >= 10 000 rpm** to wash the spin column membrane
14. Discard flow-through
15. Transfer column into a new 2 mL collection tube
16. Add **500 ul Buffer RPE** to the RNeasy spin column
17. Centrifuge for **2 min at >=13 000 rpm** to wash the spin column membrane

The long centrifugation dries the spin column membrane, ensuring that no ethanol is carried over during RNA elution. Residual ethanol may interfere with downstream reactions. After centrifugation remove the RNeasy spin column from the collection tube so that the column doesn't contact the flow-though – avoid carryover ethanol.

RNA elution

1. Place the RNeasy spin column in a new **1.5 mL collection tube**
2. Add between **30-50 ul of RNase-free water** directly to he spin column membrane

   Amount of water largely depends on the quantity of tissue used and the required RNA concentration obtained

3. Incubate **1-2 min at RT**
4. Centrifuge for **1min at >=10 000 rpm** to elute RNA

Note: Incubation of the sample for 1-2min prior to the spin is an optional step intended to improve yield. Because the subsequent enzymatic synthesis of CDNA from the RNA requires a high starting concentration of RNA (1ug or more), it is suggested to use the minimum elution volume. If desired, the elution step can be repeated to attain any residual RNA from the column

DNAse treatment

1. DNAse treat the samples with **Ambion DNAseI**
2. Add **0.1** volumes of **10xbuffer** and **1uL of DNAseI**
3. Incubate at **37°C** for **10 min**
4. Add **0.1** volumes of inactivation reagent
5. Mix gently by flicking

6. Incubate at RT for **2 min**
7. Centrifuge for **1min** at **10,000 rpm** to pellet
8. Transfer supernatant to a new tube

Quality control

1. Pipet **2ul of RNA elution** into three different 0.2 or 0.5 mL RNAse free tubes for the quantity and quality tests (Qubit, NanoDrop and Bioanalyser)
2. Store sample at **4°C for QC** or **freeze to store (-20/-80°C)**

Appendix D

Total RNA extraction protocol for sRNA sequencing

Total RNA extraction protocol for sRNA sequencing

RNA extraction: Isopropanol-chloroform protocol

Sample preparation

1. The tissue is immersed in RNAlater

    Use a wipe to remove excess and immediately add tissue to a 2 ml tube containing 1ml of TRIzol using sterile forcepsNote: 1ml of TRIzol is appropriate for 50-100mg of tissue

Homogenization

1. Homogenise the starting material using the tissue lyser (F106)

    1. Place one bead per tube with the right amount of tissue and TRIzol

    2. Homogenise at frequency 30 for 2 minutes (program 2)3. Check the bottom of the tubes and if it has cracked during the homogenisation, transfer the solution to a new tube with a bead 4. Repeat the homogenisation at frequency 30 for 2 minutes

Note: First run program 4 (10 s) on the tissue lyser to make sure the machine is well set up. The two screws turn opposite ways, and both Quiagen symbols need to face the same way. For better homogenization of the material change orientation of the Quiagen symbols after the first 2min homogenization. There is only one way the eppendorf tubes will fit the blocks – the lids have to be facing the plastic ridges on the blocks.

3. At this point the tissue can be stored in TRIzol at -80ºC overnight if necessary

RNA extraction

1. Add 200 µL chloroform
2. Shake by hand for 15 sec, incubate at room temp for 3 min.
3. Centrifuge at 4°C, 15 min, top speed.
4. Remove upper aqueous phase (approx. 600 µl) to new low-bind DNA tube.
5. Discard lower TRIzol phase appropriately
6. Add one volume (approx. 600 µl) of Chloroform, shake by hand, and repeat steps 2 to 4.
7. Add 1 µL glycogen (from Roche) as a carrier and 500 µL isopropanol

Note: If samples were “cloudy” upon suspension in isopropanol because of high salt concentration coming from RNAlater I increased the dilution factor: samples were eventually suspended in a 1900 µL solution composed of (i) isopropanol and (ii) DEPC water + aqueous phase (1:1 ratio between i and ii) in 2 ml low DNA binding Eppendorf tubes.

8. Mix and incubate at -20°C overnight
9. Centrifuge at 4°C, 30 min, top speed. Expect relatively small RNA pellet.
10. Remove supernatant and add 800 µl of 70% EtOH. Invert a few times to let the pellet swim in the EtOH solution and centrifuge at 4°C, 5 min, top speed.
11. Repeat step 9 once.
12. Remove supernatant with a 1 ml tip, spin briefly, carefully remove the rest of the liquid with a 10µl tip.
13. Air dry for 2 min at room temp (just time for the pellets to become translucent)
14. Re-suspend in 20µl of Illumina pure H_2O
15. Use 1 µl to Nanodrop the samples
16. Store at -80°C